Fgenes

Pattern based human gene structure prediction (multiple genes, both chains).

Method description:

Algorithm based on pattern recognition of different types of exons, promoters and polyA signals. Optimal combination of these features is then found by dynamic programming and a set of gene models is constructed along a given sequence.

Fgenes output:

G - predicted gene number, starting from start of sequence;

Str - DNA strand (+ for direct and - for complementary strands);

Feature - type of coding sequence: CDSf - First (Starting with Start codon), CDSi - internal (internal exon), CDSl - last coding segment, ending with stop codon);

TSS - position of transcription start;

TATA – position of TATA-box;

wTATA – Discriminant function score for TATA box;

TSS - Positions of transcription start (TATA-box position and score);

Start and End - Position of the Feature;

Weight - Discriminant function score for the feature;

ORF - start/end positions of ORF where the first complete codon starts and the last codon ends.

```
FGENES 1.5 Prediction of multiple genes in genomic DNA
Time: 171940.7 Date: 20001003
Seq name: > HUMHBB       73308 bp    DNA              PRI        20-JAN-1
Length of sequence:   73308 GC content: 0.39 Zone: 1
Number of predicted genes:   9 In +chain:   7 In -chain:   2
Number of predicted exons:  23 In +chain:  19 In -chain:   4
Positions of predicted genes and exons:
 G Str Feature  Start         End    Weight   ORF-start ORF-end

 1 -   1 CDSi    5978 -      6039    1.69     5978 -     6037
 1 -   2 CDSf    6314 -      6365    1.40     6315 -     6365

 2 -   1 CDSl   13709 -     13807    1.84    13712 -    13807
 2 -   2 CDSf   14781 -     14855    1.62    14781 -    14855

 3 +     TSS    19488              5.83 TATA  19457 wTATA   19.85 LDF    0.81
 3 +   1 CDSf   19541 -     19632   11.08    19541 -    19630
 3 +   2 CDSi   19755 -     19977    6.20    19756 -    19977
 3 +   3 CDSl   20833 -     20961    5.95    20833 -    20958
 3 +     PolA   21055              2.08

 4 +     TSS    34478              4.98 TATA  34447 wTATA   19.21 LDF    0.91
 4 +   1 CDSf   34531 -     34622    8.82    34531 -    34620
 4 +   2 CDSi   34745 -     34967    5.96    34746 -    34967
 4 +   3 CDSl   35854 -     35982    6.30    35854 -    35979
 4 +     PolA   36043              2.68

 5 +     TSS    39412              5.00 TATA  39383 wTATA   19.21 LDF    0.93
 5 +   1 CDSf   39467 -     39558    8.82    39467 -    39556
 5 +   2 CDSi   39681 -     39903    5.96    39682 -    39903
 5 +   3 CDSl   40770 -     40898    6.17    40770 -    40895
 5 +     PolA   40959              2.78

 6 +   1 CDSf   45995 -     46151    3.09    45995 -    46150
 6 +   2 CDSl   46997 -     47100    2.32    46999 -    47097
 6 +     PolA   47243              2.75

 7 +   1 CDSf   54790 -     54881    8.97    54790 -    54879
 7 +   2 CDSi   55010 -     55232    5.60    55011 -    55232
```

```
 7 +    3 CDSl    56131 -    56259    5.05    56131 -    56256
 7 +      PolA    56365             1.07

 8 +    1 CDSf    62187 -    62278    9.72    62187 -    62276
 8 +    2 CDSi    62409 -    62631    6.64    62410 -    62631
 8 +    3 CDSl    63482 -    63610    6.56    63482 -    63607
 8 +      PolA    63718             4.72

 9 +    1 CDSf    68183 -    68290    2.50    68183 -    68290
 9 +    2 CDSl    70703 -    70819    1.10    70703 -    70816
 9 +      PolA    70905             4.71
```

Predicted proteins:
```
>FGENES 1.5 > HUMHBB      7   1 Multiexon gene     5978 -     6365      38 a Ch-
MVCNCGLDHNFQSPRSKTCAFNKLIYTTSTLGSSSINE
>FGENES 1.5 > HUMHBB      7   2 Multiexon gene    13709 -    14855      57 a Ch-
MCSHHLASNCCFRSVPLPHLSRSLQEFVLKVNFHNRKLIEAKASVKERNISSKPLCC
>FGENES 1.5 > HUMHBB      7   3 Multiexon gene    19541 -    20961     147 a Ch+
MVHFTAEEKAAVTSLWSKMNVEEAGGEALGRLLVVYPWTQRFFDSFGNLSSPSAILGNPK
VKAHGKKVLTSFGDAIKNMDNLKPAFAKLSELHCDKLHVDPENFKLLGNVMVIILATHFG
KEFTPEVQAAWQKLVSAVAIALAHKYH
>FGENES 1.5 > HUMHBB      7   4 Multiexon gene    34531 -    35982     147 a Ch+
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTGVASALSSRYH
>FGENES 1.5 > HUMHBB      7   5 Multiexon gene    39467 -    40898     147 a Ch+
MGHFTEEDKATITSLWGKVNVEDAGGETLGRLLVVYPWTQRFFDSFGNLSSASAIMGNPK
VKAHGKKVLTSLGDAIKHLDDLKGTFAQLSELHCDKLHVDPENFKLLGNVLVTVLAIHFG
KEFTPEVQASWQKMVTAVASALSSRYH
>FGENES 1.5 > HUMHBB      7   6 Multiexon gene    45995 -    47100      86 a Ch+
MGNPKVKAHGKKVLISFGKAVMLTDDLKGTFATLSDLHCNKLHVDPENFLVSTLRQRDID
CFGNPLQRGFYPTDTGFLAVTNKCCG
>FGENES 1.5 > HUMHBB      7   7 Multiexon gene    54790 -    56259     147 a Ch+
MVHLTPEEKTAVNALWGKVNVDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPENFRLLGNVLVCVLARNFG
KEFTPQMQAAYQKVVAGVANALAHKYH
>FGENES 1.5 > HUMHBB      7   8 Multiexon gene    62187 -    63610     147 a Ch+
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
>FGENES 1.5 > HUMHBB      7   9 Multiexon gene    68183 -    70819      74 a Ch+
MEQSWAENDFDELREEGFRRSNYSKLKEEVRTNGKEASIILIPKPDRDTTKKENVTPISL
MNIDAKILNKILAN
```