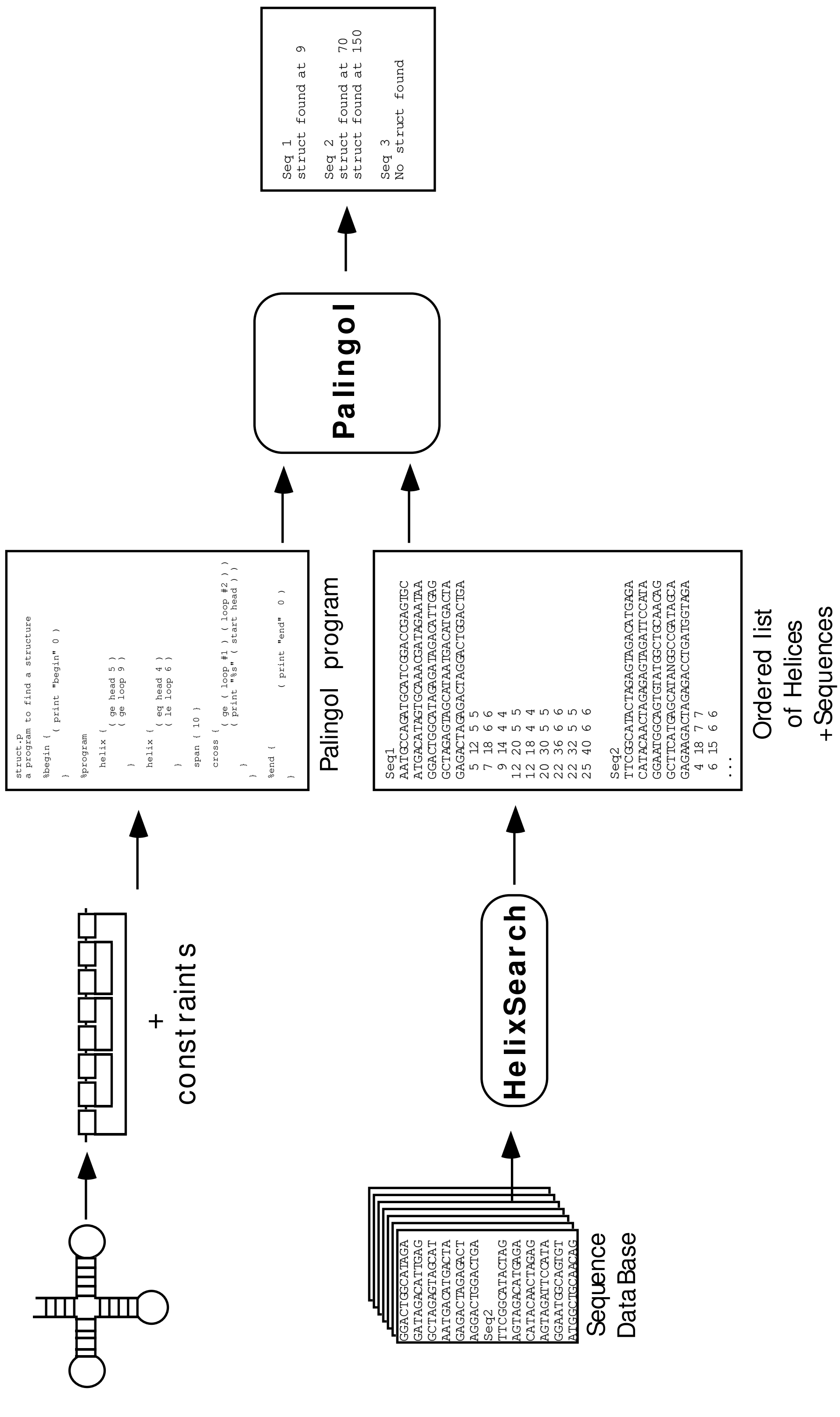

Figure 1

The overall process of searching a structure in a sequence database. The process starts with the description of the required structure as a set of elementary helices together with constraints. Local constraints act on each helix, whereas global constraints specify arbitrarily complex relationships between them. These constraints are written using the Palingol syntax, giving rise to a Palingol program. The elementary helices are searched by an external program (HelixSearch) that can be supplied by the user (a default HelixSearch program is however supplied with Palingol). Then the Palingol interpreter reads the program and, by evaluating the constraints on the list of helices, searches for all valid subsets of helices.

Table 1
Searching for IRE in 4 test sets of vertebrate sequences.
Total: total number of nucleotides scanned ;
Number of positive sequences meeting increasingly stringent requirements:

MOTIF: consensus motif, CNNNNNCAGTGH ;
TOP: whole top helix, i.e. MOTIF and an helix of 4 or 5 base pairings ;
IRE: recognized as true IRE: TOP and bottom helix;
Time: total time required on a SPARC20 workstation to scan the set with the complete IRE program.

Table 2
The 22 IREs and TOPs found in Ferritin related sequences.
(a): 19 found IREs.
(b): 3 missed IREs. Mismatching and bulgy nucleotides are underlined. Missing nucleotides are replaced by dots.

Table 3
Comparison of results found by tRNAscan and the Palingol program trna.p.
Numbers in parenthesis are percentages computed using the method described in Fichant and Burks 1991). Set of sequences: Prokaryotic sequences in EMBL43, with size < 1000 bases: 8998 sequences, totaling 3,625,000 nucleotides. This set contains 678 annotated tRNA sequences on the direct strand, without intron. Unannotated tRNA sequences are considered as not being tRNAs, thus increasing the apparent number of false positives. Total CPU time (SPARC670 workstation): tRNAscan: 5 min.; trna.p: 12 min. (preprocessing by Palamou: 13 min.).

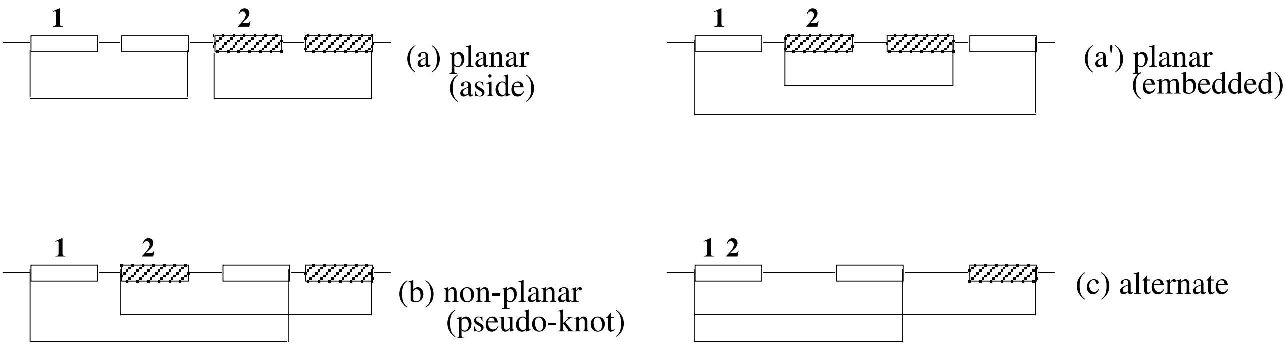

Figure 2
Some examples of structures that can be represented in Palingol.
(**a**) and (**a'**) are pure secondary structures; (**b**) is a non-nested (tertiary) structure and (**c**) is an 'alternate' structure that may correspond to an equilibrium (like in bacterial attenuators for instance).

(a)

```
        helix {
              ... local constraints
                    of helix 1
        }
        helix {                              | helices part

              ... local constraints
                    of helix 2
        }

        ...

        span {
              ... span constraints           | span part
        }

        cross {
              ... global constraints         | cross part
                    and output

        }

}
```

(b)

```
        span {
              %keyword value                 | (# helices - 1)
              ...                            | lines
              %keyword value
        }
```

Figure 3
A Palingol program layout. The *main* part is composed of three subparts respectively called *helix*, *span* and *cross* part. The *helix* part(s) specify local constraints acting on each elementary helix, the *span* part is optional and is used to prune the Palingol engine's search; the *cross* part specifies global constraints acting across different helices.

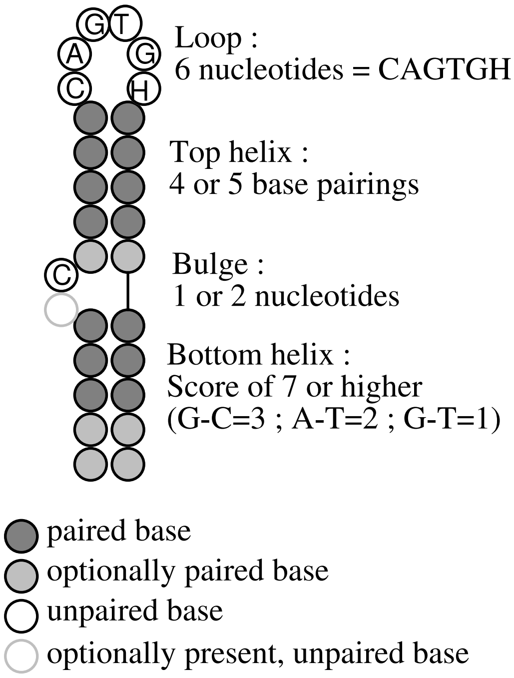

Figure 4
The traditional representation and description of an Iron Responsive Element (IRE).

Step	Constraint	Actual
1. TΨC signal	≥ 2/4 in GNTCNNNNC score ≥ 0.4 (SG++ if ≥ 3/4)	≥ 43 from 1st nt in aa head ≤ 23 from last nt in aa tail
2. TΨC arm	loop = 7 helix = 4 or 5 (SG++ if helix ≥ 5)	helix ≥ 4 head at base 2 or 3 of signal (head+loop+tail) ≤ 17
3. D signal	≥2 / 3 in TNGNNNA score ≥ 0.4 (SG++ if = 3/3)	starts 37-60 nt upstr from TΨC and ≥ 8 nt from 1st nt in aa arm head
4. D arm	helix = 2 or 3 (SG++ if helix ≥ 3)	helix ≥ 2 15 ≤ (head+loop+tail) ≤ 19 head starts at 3rd nt of D signal.
5. Amino-acyl arm	helix = 6 or 7 (SG++ if helix ≥ 7)	helix ≥ 6 head starts 7 nt upstr. from D signal tail ends 24 nt downst. from TΨC signal
TEST SG ≥ 4		
6. Anticodon arm	loop = 7 helix = 4 or 5	helix ≥ 4 starts 2 nt from end of D arm (head+loop+tail) ≤ 17
7. 5' of anticodon	5' base = T (SG++ if yes)	T 6 nt from start of AC zone
TEST SG ≥ 5		

Figure 6
Summary of the tRNAscan algorithm and rewriting as structural constraints. The *Step* column indicates the seven steps of the Fichant's algorithm. The *Constraint* column indicates the corresponding Palingol constraints. The *Actual* column indicates which constraints were actually applied when ambiguities arose in the original paper or when the original specifications had to be slightly modified.

```
# ire.p
# a program to look for Iron Responsive Elements.

# Helix Search Parameters:
# - helix 4 or 5; score >= 7; loop = 6
# - helix >=3; score >=7; loop = 17 or 18
# helix score computed with: GC=3 ; AT=2 ; GT=1

%start {
   (print "Palingol search for IREs\n" 0)
   (bpcompile "GC 3 AT 2 GT 1")
}

%program {

# Bottom helix
 helix {
   (or (eq loop 17) (eq loop 18))
   (ge (score head tail) 7)
 }

# Top helix
 helix {
   (eq loop 6)
   (set $zone (sstr (seq fullseq)
              (sub (start loop) 6) 12))
   (patsearch $zone
              "CNNNNNCAGTGH" 1 0 false)
 }

 span { %end_head 4 }

 cross {

# Bulge: 0 (resp. 1) nt if bottom loop = 17 (resp. 18) nt
  (set $bul (sub loop #1 16))

# Mismatch: 0 (resp. 1) nt if 5 (resp. 4) in top helix
  (set $mis (sub 5 head #2))

# Heads separated by buldge and mismatch
  (eq (start head #2) (add (end head #1)
                (add (add $bul $mis) 1)))

# Tails separated by mismatch
  (eq (start tail #1) (add (end tail #2)
                           (add $mis 1)))

# Print results
  (set $pos (start head #1))
  (set $size (add (add (head #1)
                  (loop #1)) (tail #1)))

  (print "%s " (seq seqname))
  (print "IRE at %d " $pos)
  (print " %s\n" (sstr (seq fullseq)
                          $pos $size))
 }

# End of main program
}

%end {
  (print "End of Search\n" 0)
}
```

Figure 5

A Palingol sample program to search for IREs. Lines starting with a '#' are considered as comments by the interpreter.

Set of sequences	Total (nt)	MOTIF	TOP	IRE	Time
Ferritin	59 943	51	22	19	28 sec
5' non coding	6 971 415	1705	66	34	58 min
3' non coding	12 010 172	3199	98	47	1 h 43 min
introns	9 129 148	2266	43	16	1 h 17 min

(a)

Mnemonic	size	IRE position	Complete IRE sequence
GGFERH	7126	1306 – 1336	GTTCCTG C GTCAA CAGTGC TTGGA CGGAACC
HSAFLP1	457	2 – 28	TCTTG C TTCAA CAGTGT TTGGA CGGAA
HSFERG1	512	207 – 239	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
HSFERHX	2083	949 – 981	GTTTCCTG C TTCAG CAGTGC TTGGA CGGAACCC
HSFERP1	1052	118 – 150	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
HSFERP2	1036	118 – 150	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
HSFERRITH	1198	29 – 61	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
HSFHC122	214	30 – 62	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
MMFERHC	2790	229 – 261	GTTTCCTG C TTCAA CAGTGC TTGAA CGGAACCC
MMFERHG	3322	967 – 999	GTTTCCTG C TTCAA CAGTGC TTGAA CGGAACCC
MMFERRH	1109	967 – 999	GTTTCCTG C TTCAA CAGTGC TTGAA CGGAACCC
OCFERL5	98	31 – 61	TGTCTTG C TTCAA CAGTGT TTGAA CGGAACA
RCFERH	853	24 – 54	GTTCTTG C TTCAA CAGTGT TTGAA CGGAACC
RNFERA1	563	205 – 237	GTTTCCTG C TTCAA CAGTGC TTGAA CGGAACCC
RNFERL1	2365	400 – 430	TATCTTG C TTCAA CAGTGT TTGGA CGGAACA
RNFERUTR	66	20 – 50	TATCTTG C TTCAA CAGTGT TTGGA CGGAACA
SSFE	821	12 – 44	GTTTCCTG C TTCAA CAGTGC TTGGA CGGAACCC
XLFERHSU	868	7 – 37	GTTCTTG C TTCAA CAGTGT TTGAA CGGAACC
XLFERRIT	1036	166 – 198	AGTTCTTG C TTCAA CAGTGT TTGAA CGGAACCT

(b)

Mnemonic	size	TOP position	TOP sequence and surrounding bases
HSAFL12	754	141 – 154	gtctcttg c TTCAA CAGTGT TTGAC gaacagat
MMFERLSUB	3015	1198 – 1213	tgtacttg c TTCAA CAGTGT TTGAA cggaaca
MMFERLA	889	6 – 21	.....cttg c TTCAA CAGTGT TTGAA cggaacaga

	Total	True positive	False positive	False negative
tRNAscan	651	614	37 (0.079%)	64 (10.42%)
trna.p (Palingol)	645	615	30 (0.064%)	63 (10.24%)
tRNAscan **not** trna.p	29	19	10	20
trna.p **not** tRNAscan	23	20	3	19