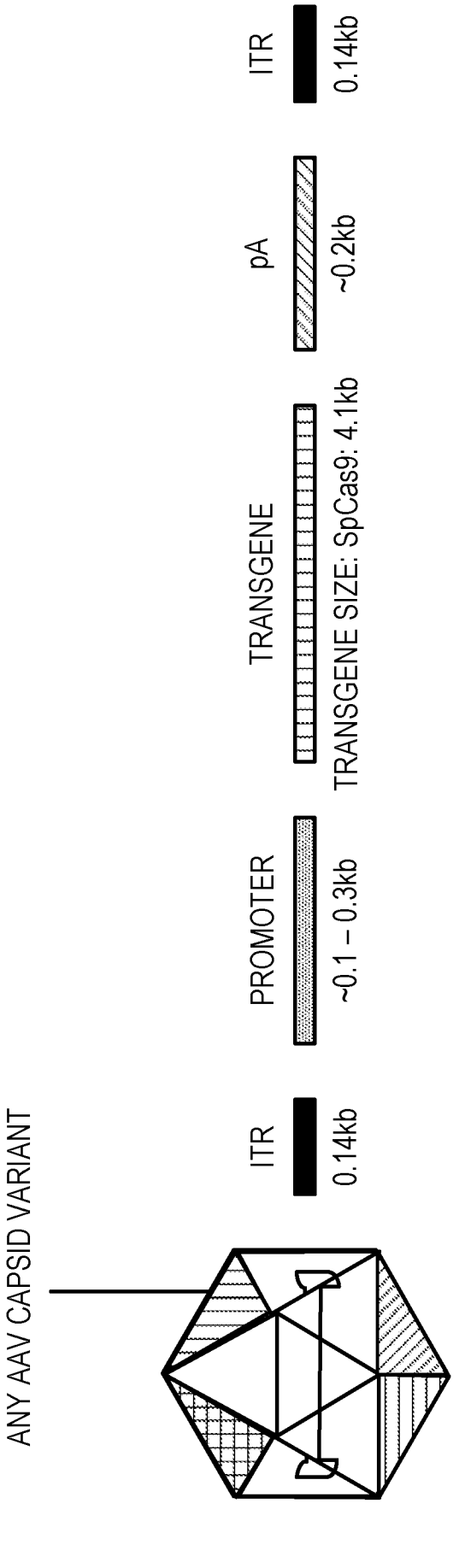

ANY AAV CAPSID VARIANT
ITR
0.14kb
PROMOTER
~0.1 – 0.3kb
TRANSGENE
TRANSGENE SIZE: SpCas9: 4.1kb
pA
~0.2kb
ITR
0.14kb
~4.7kb PACKAGING CAPACITY

FIG. 1

FIG. 2

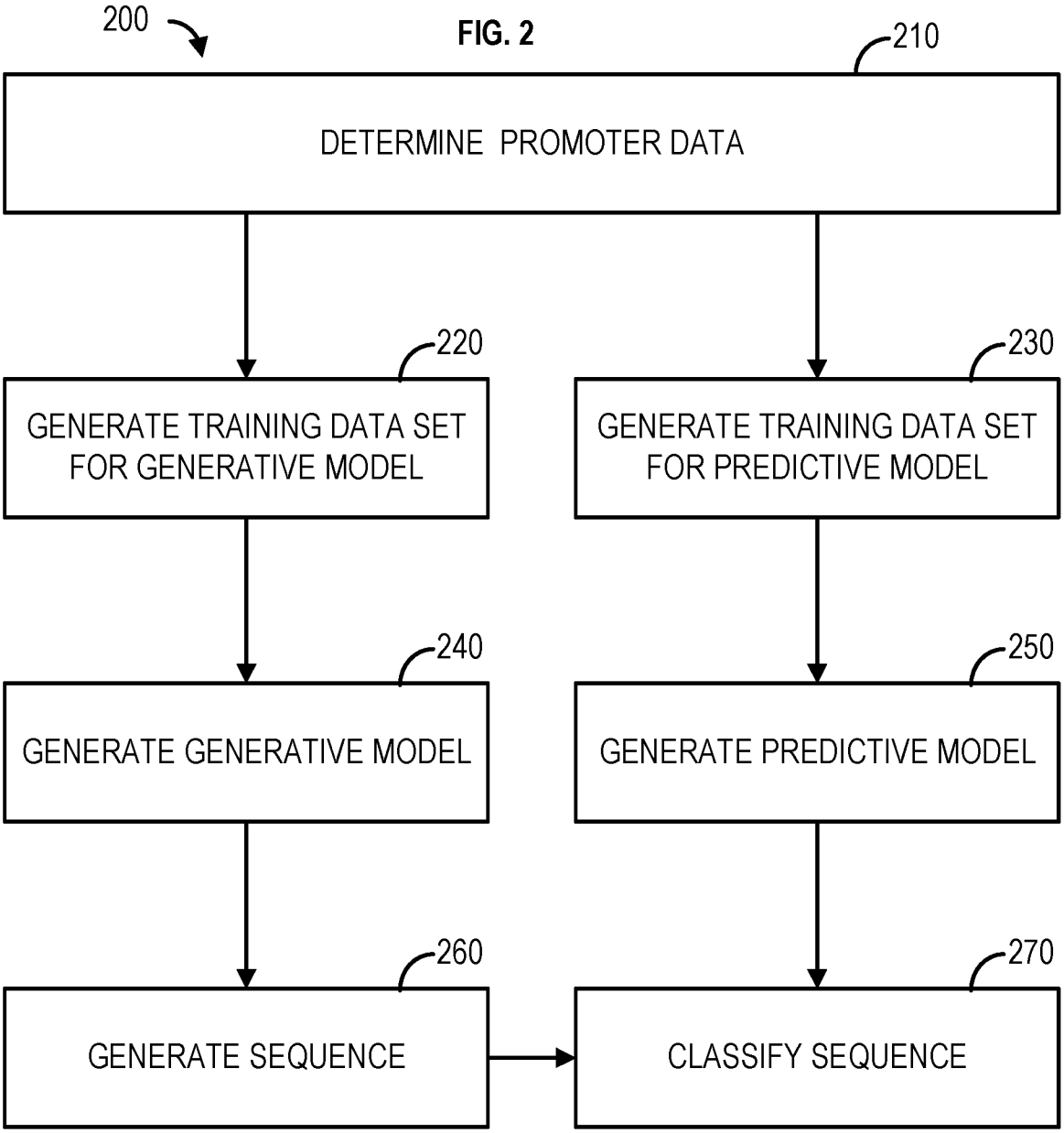

200
210
DETERMINE PROMOTER DATA
220
GENERATE TRAINING DATA SET FOR GENERATIVE MODEL
230
GENERATE TRAINING DATA SET FOR PREDICTIVE MODEL
240
GENERATE GENERATIVE MODEL
250
GENERATE PREDICTIVE MODEL
260
GENERATE SEQUENCE
270
CLASSIFY SEQUENCE

FIG. 3

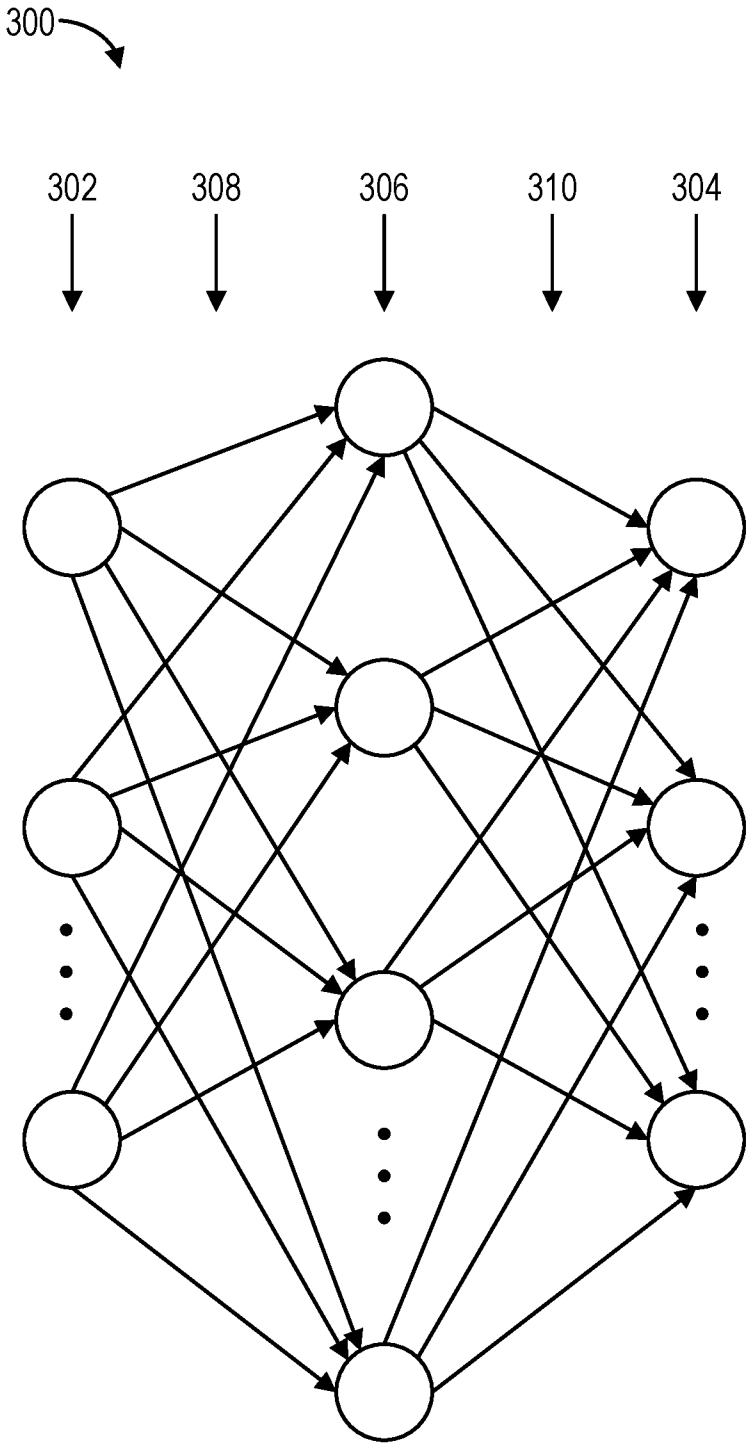


FIG. 4A

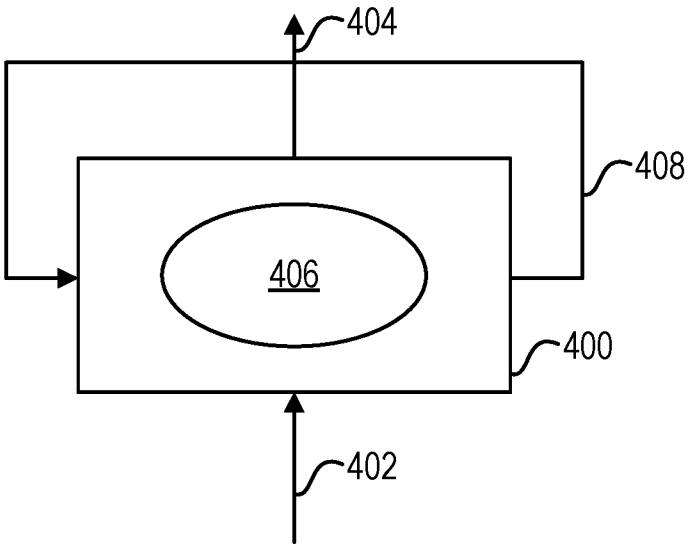
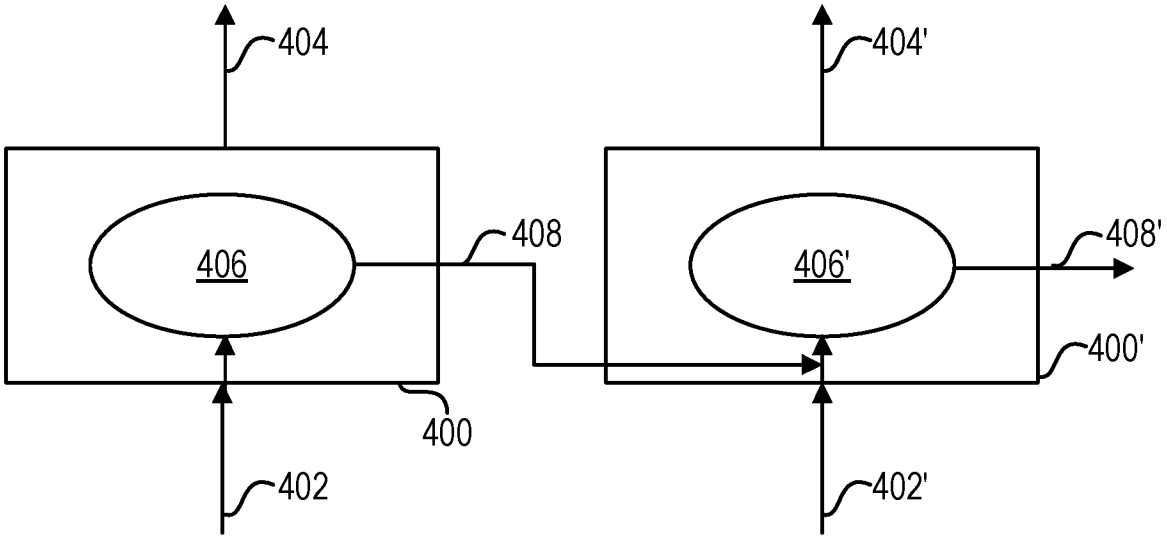


FIG. 4B



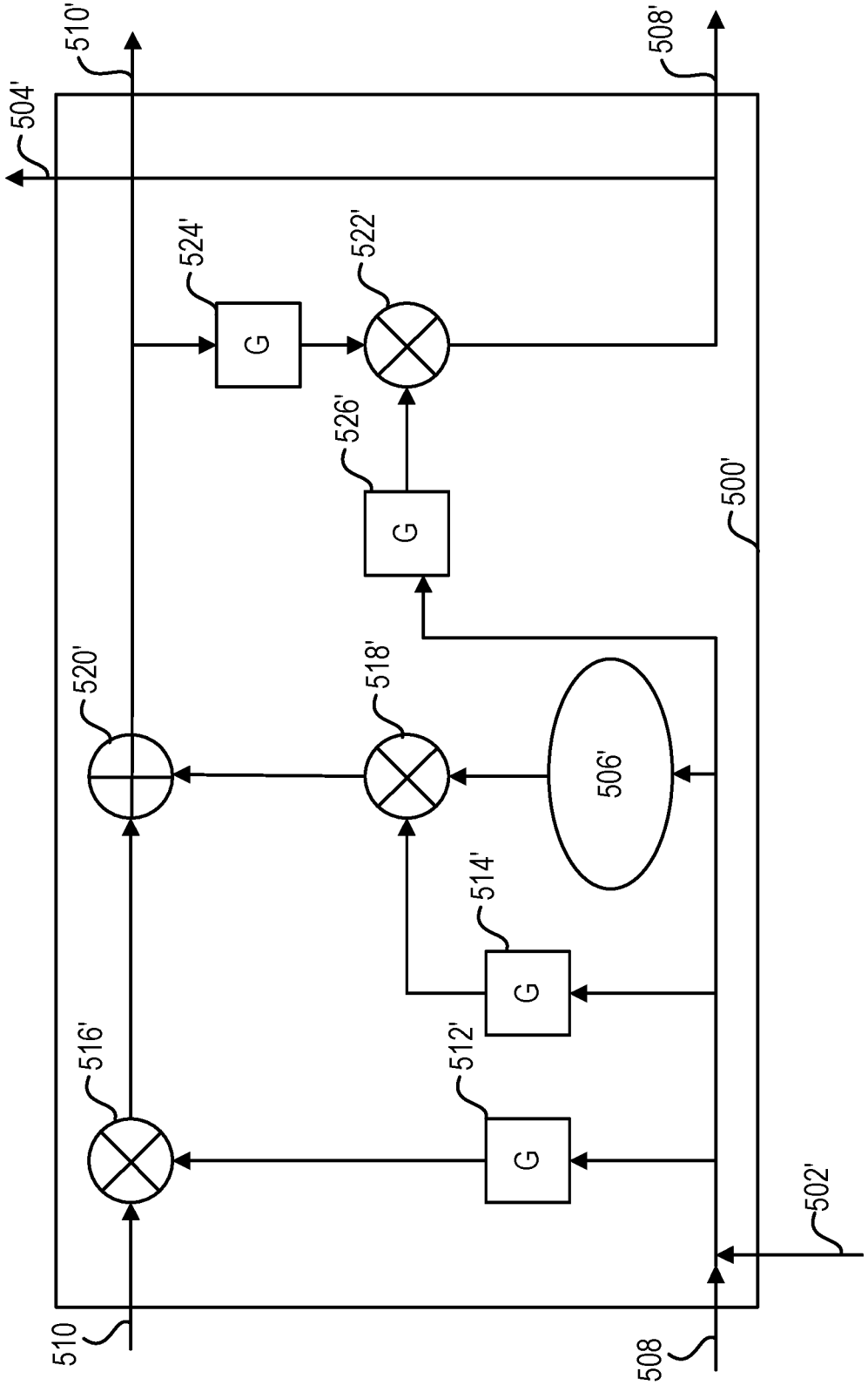

504'
510'
508'
524'
522'
G
526'
G
500'
520'
518'
506'
514'
G
516'
512'
G
502'
510
508

FIG. 5

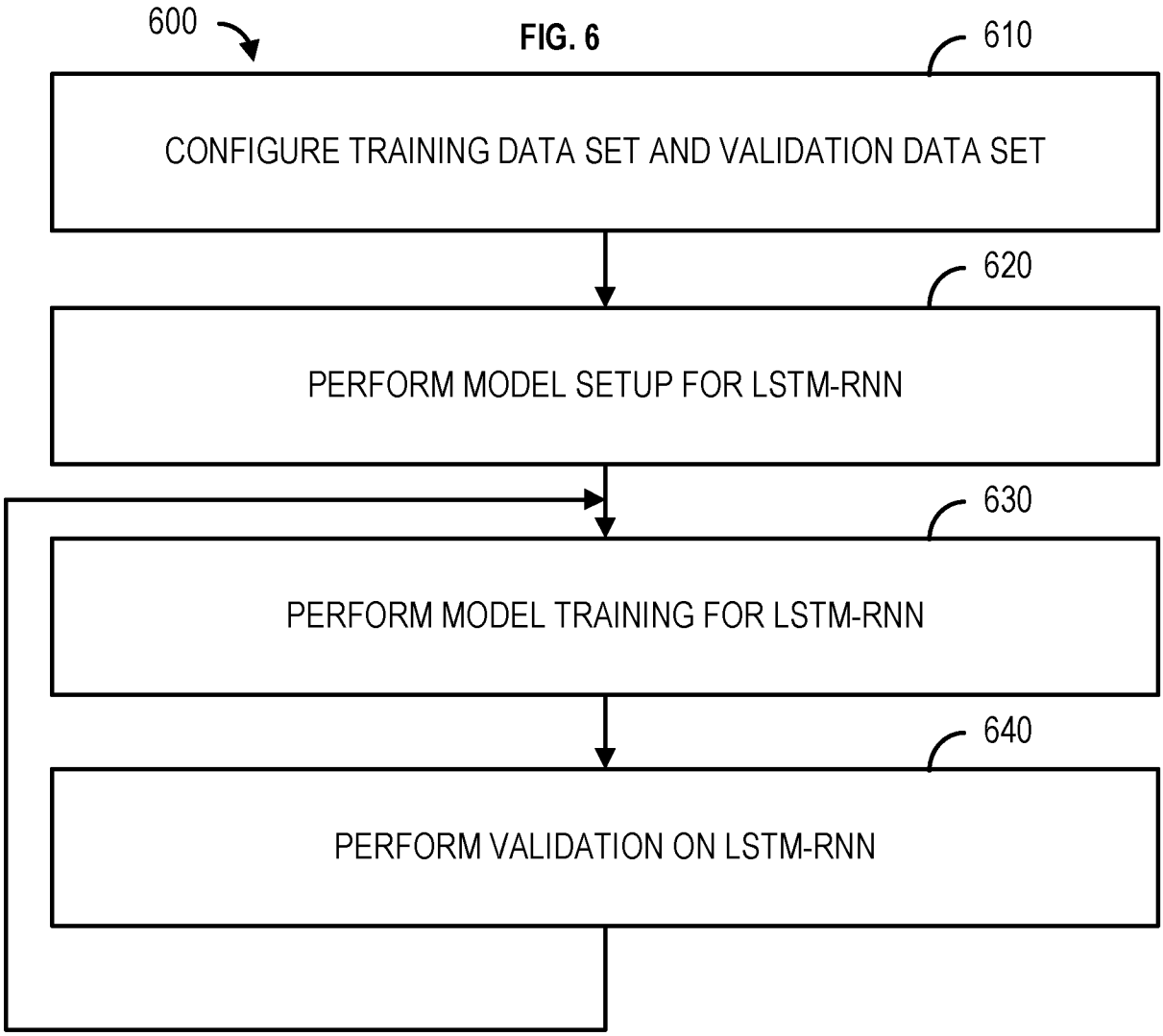

600
610
CONFIGURE TRAINING DATA SET AND VALIDATION DATA SET
620
PERFORM MODEL SETUP FOR LSTM-RNN
630
PERFORM MODEL TRAINING FOR LSTM-RNN
640
PERFORM VALIDATION ON LSTM-RNN

FIG. 6

FIG. 7

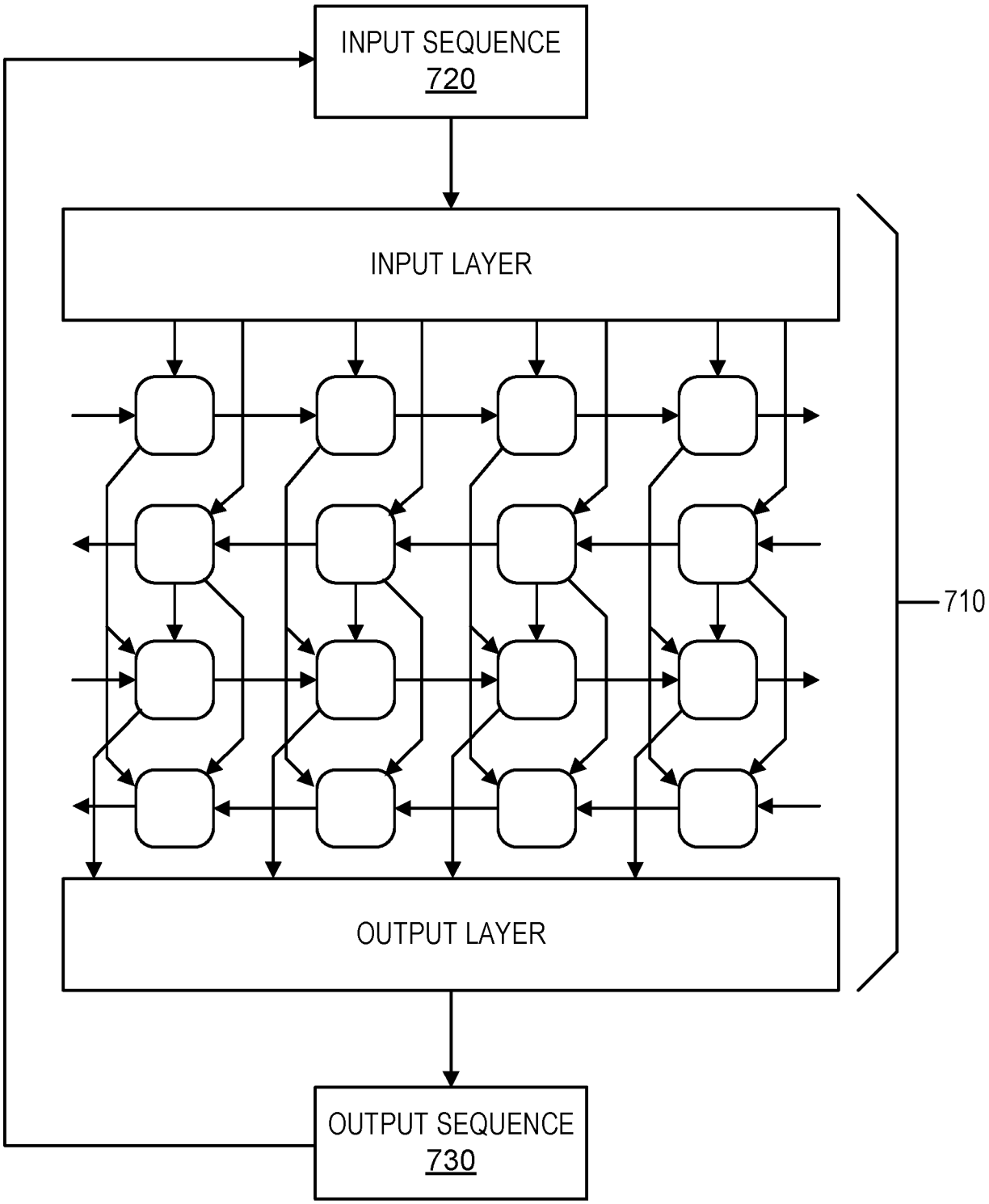
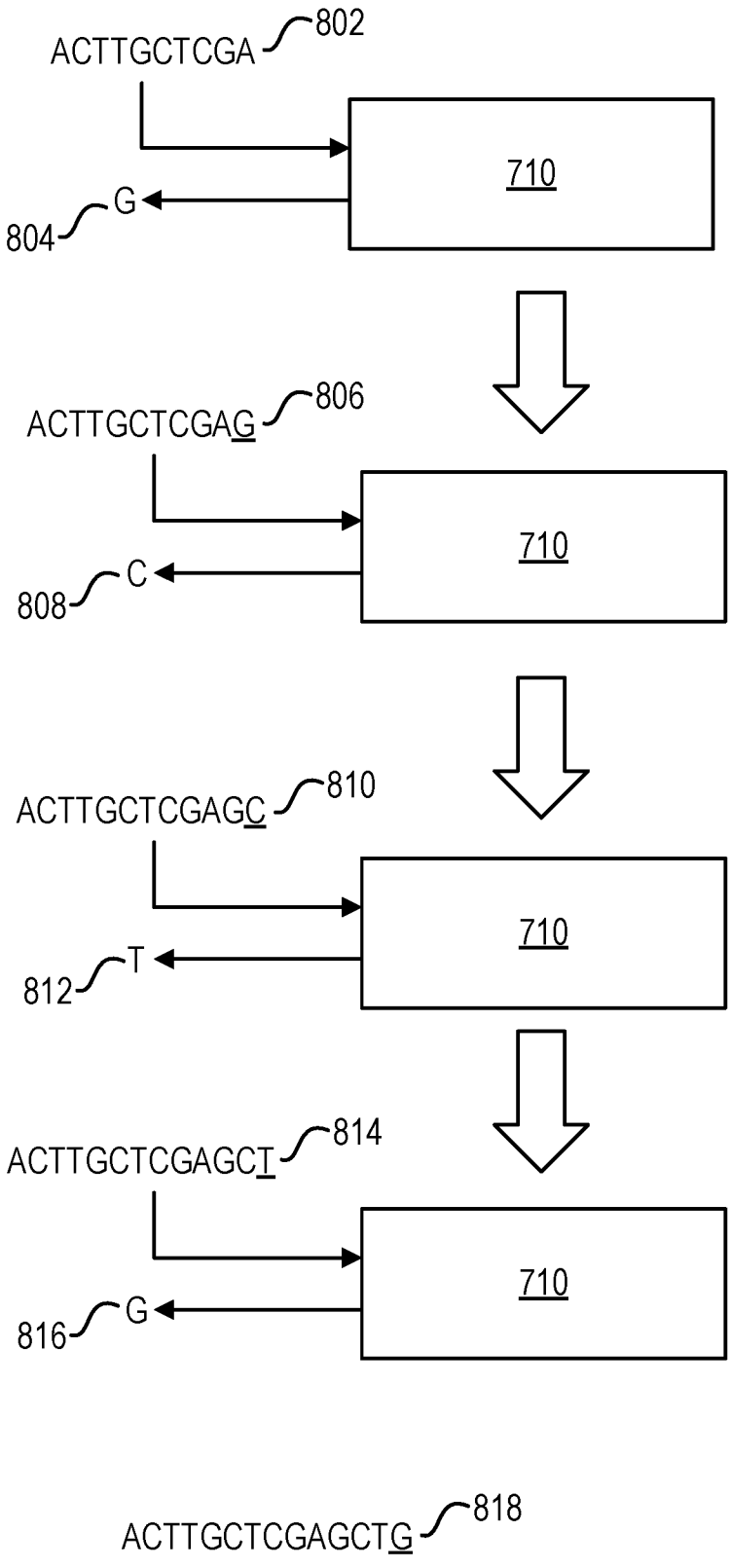

INPUT SEQUENCE
720
INPUT LAYER
OUTPUT LAYER
OUTPUT SEQUENCE
730
710

FIG. 8


ACTTGCTCGA
802
710
G
804
ACTTGCTCGAG
806
710
C
808
ACTTGCTCGAGC
810
710
T
812
ACTTGCTCGAGCT
814
710
G
816
ACTTGCTCGAGCTG
818

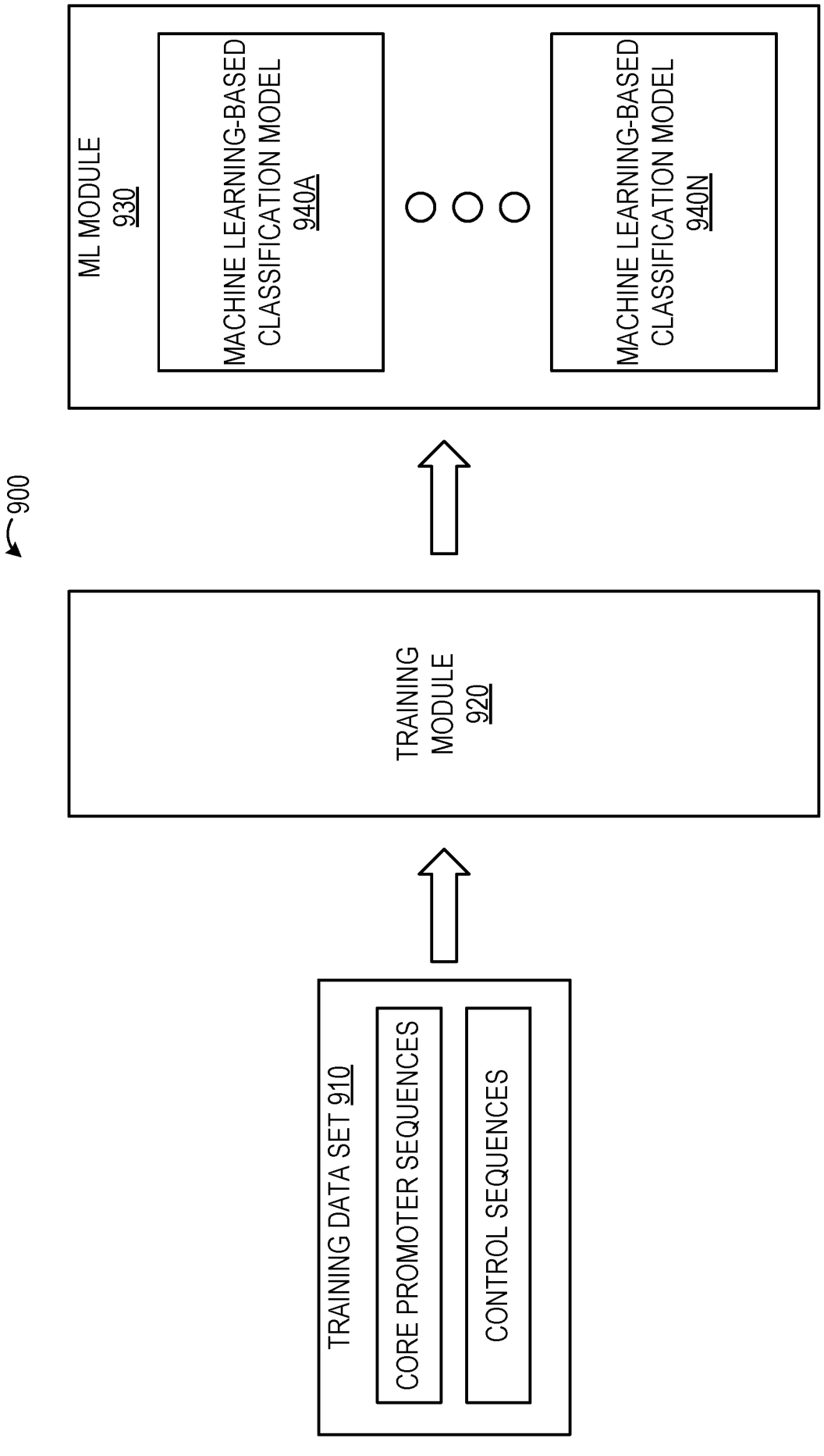

900
TRAINING DATA SET 910
CORE PROMOTER SEQUENCES
CONTROL SEQUENCES
TRAINING MODULE 920
ML MODULE 930
MACHINE LEARNING-BASED CLASSIFICATION MODEL 940A
MACHINE LEARNING-BASED CLASSIFICATION MODEL 940N

FIG. 9

FIG. 10

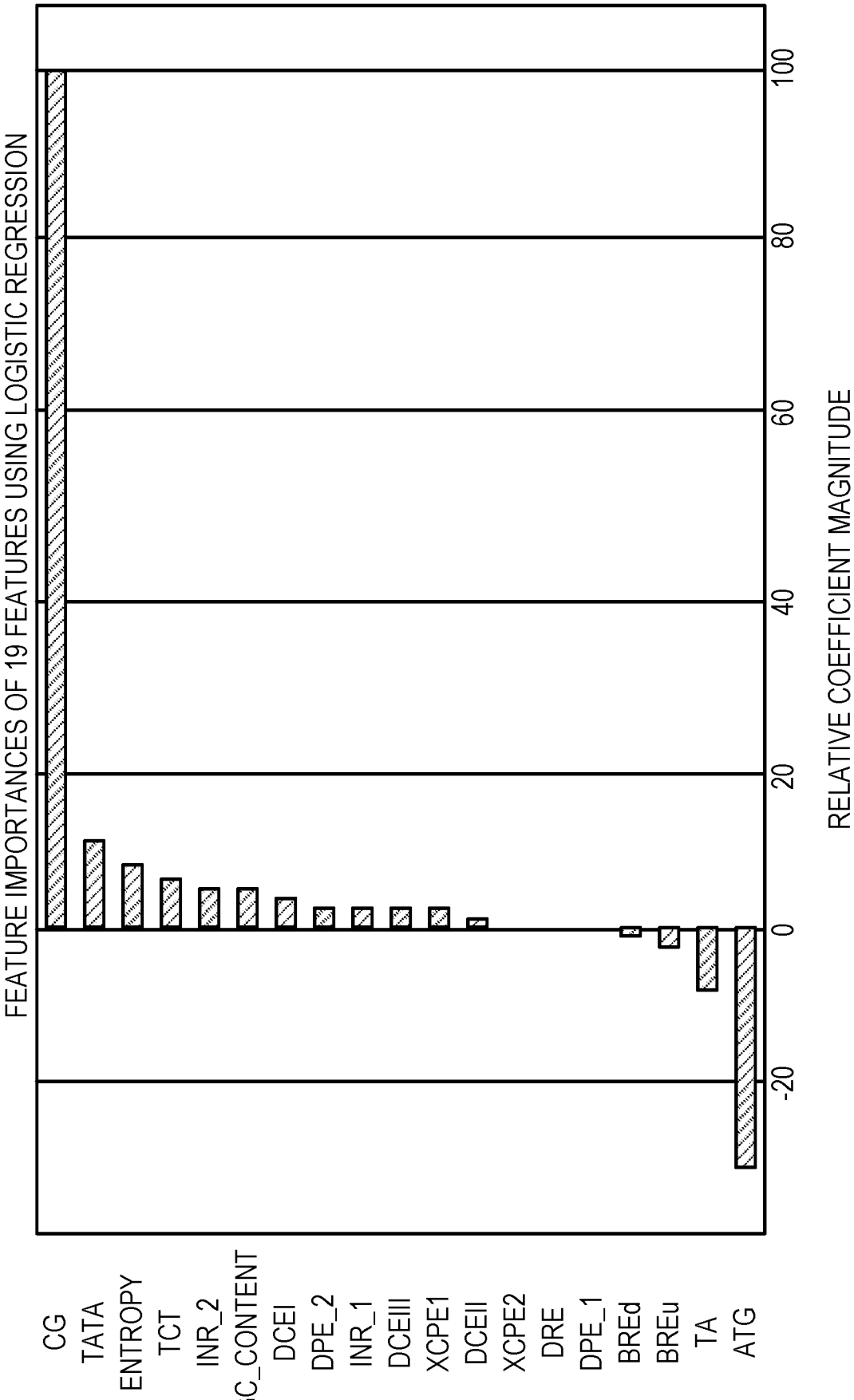
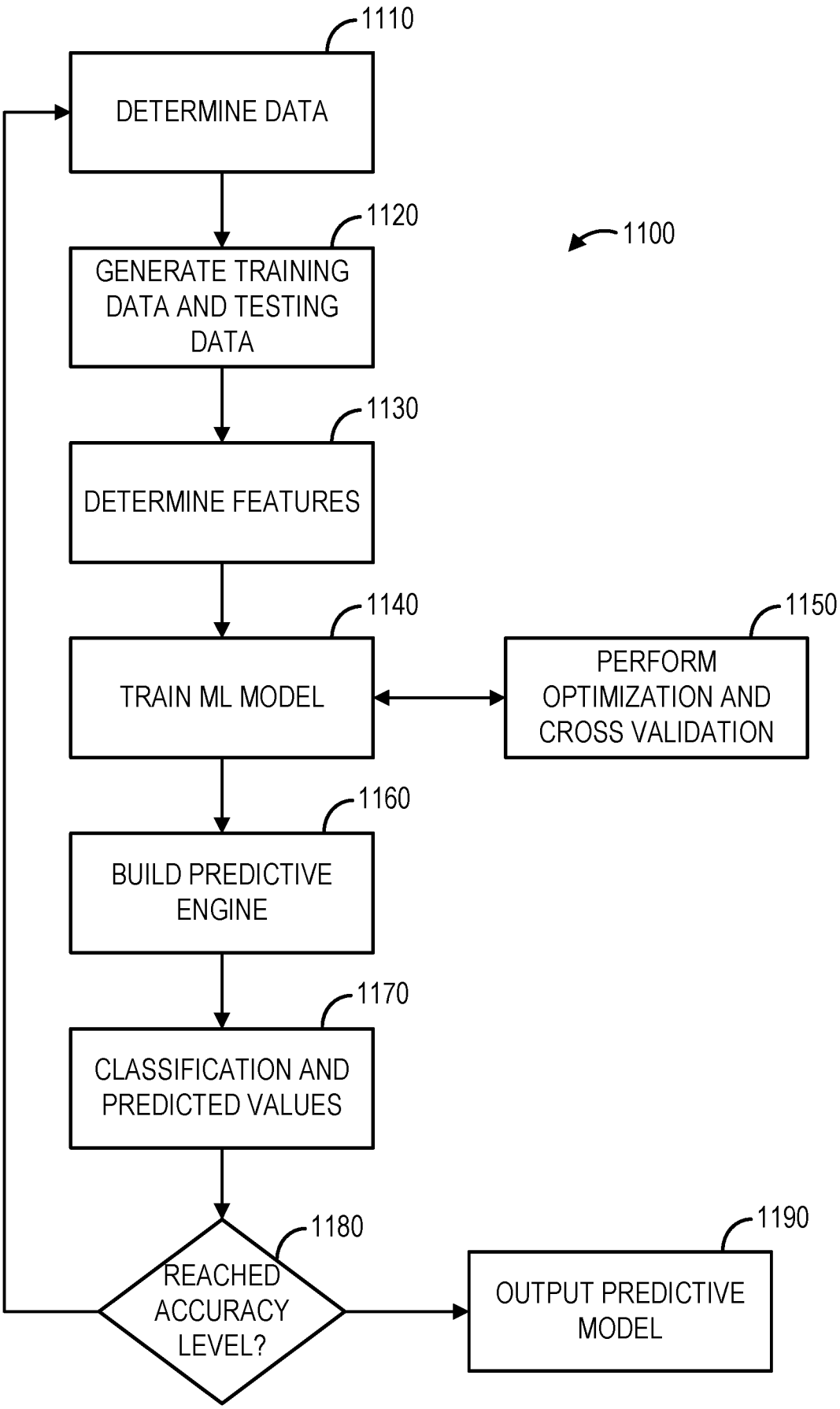

FEATURE IMPORTANCES OF 19 FEATURES USING LOGISTIC REGRESSION
RELATIVE COEFFICIENT MAGNITUDE
100
80
60
40
20
0
-20
CG
TATA
ENTROPY
TCT
INR_2
GC_CONTENT
DCEI
DPE_2
INR_1
DCEIII
XCPE1
DCEII
XCPE2
DRE
DPE_1
BREd
BREu
TA
ATG

FIG. 11


1110
DETERMINE DATA
1120
GENERATE TRAINING DATA AND TESTING DATA
1100
1130
DETERMINE FEATURES
1140
TRAIN ML MODEL
1150
PERFORM OPTIMIZATION AND CROSS VALIDATION
1160
BUILD PREDICTIVE ENGINE
1170
CLASSIFICATION AND PREDICTED VALUES
1180
REACHED ACCURACY LEVEL?
1190
OUTPUT PREDICTIVE MODEL

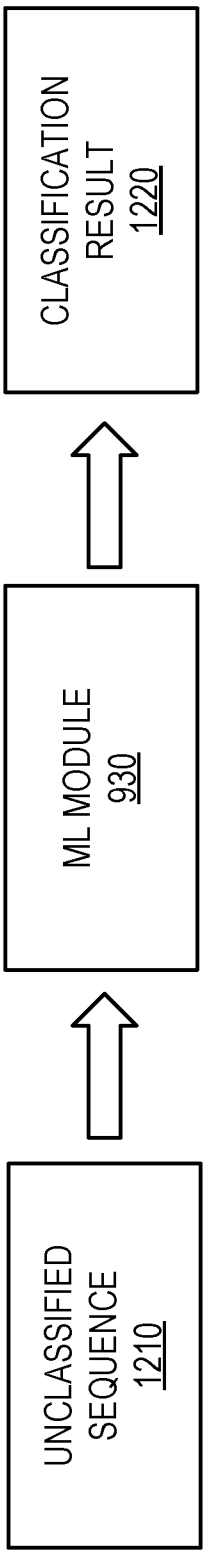

UNCLASSIFIED SEQUENCE 1210
ML MODULE 930
CLASSIFICATION RESULT 1220

FIG. 12

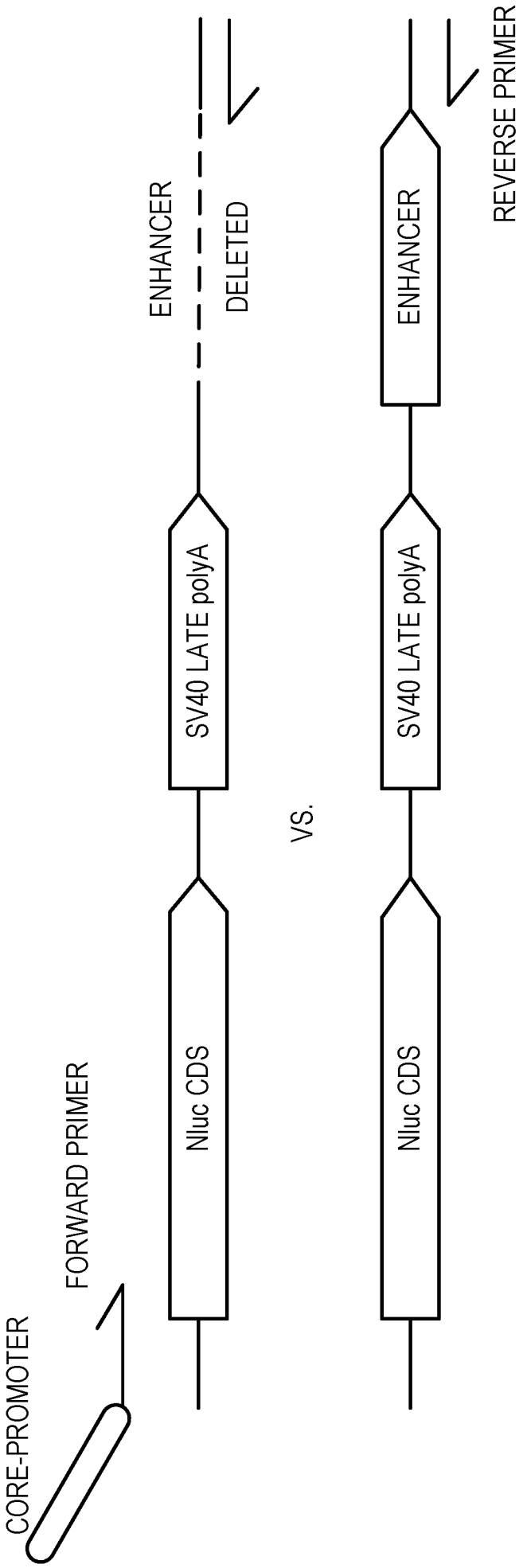

CORE-PROMOTER
FORWARD PRIMER
Nluc CDS
SV40 LATE polyA
ENHANCER
DELETED
VS.
Nluc CDS
SV40 LATE polyA
ENHANCER
REVERSE PRIMER

FIG. 13

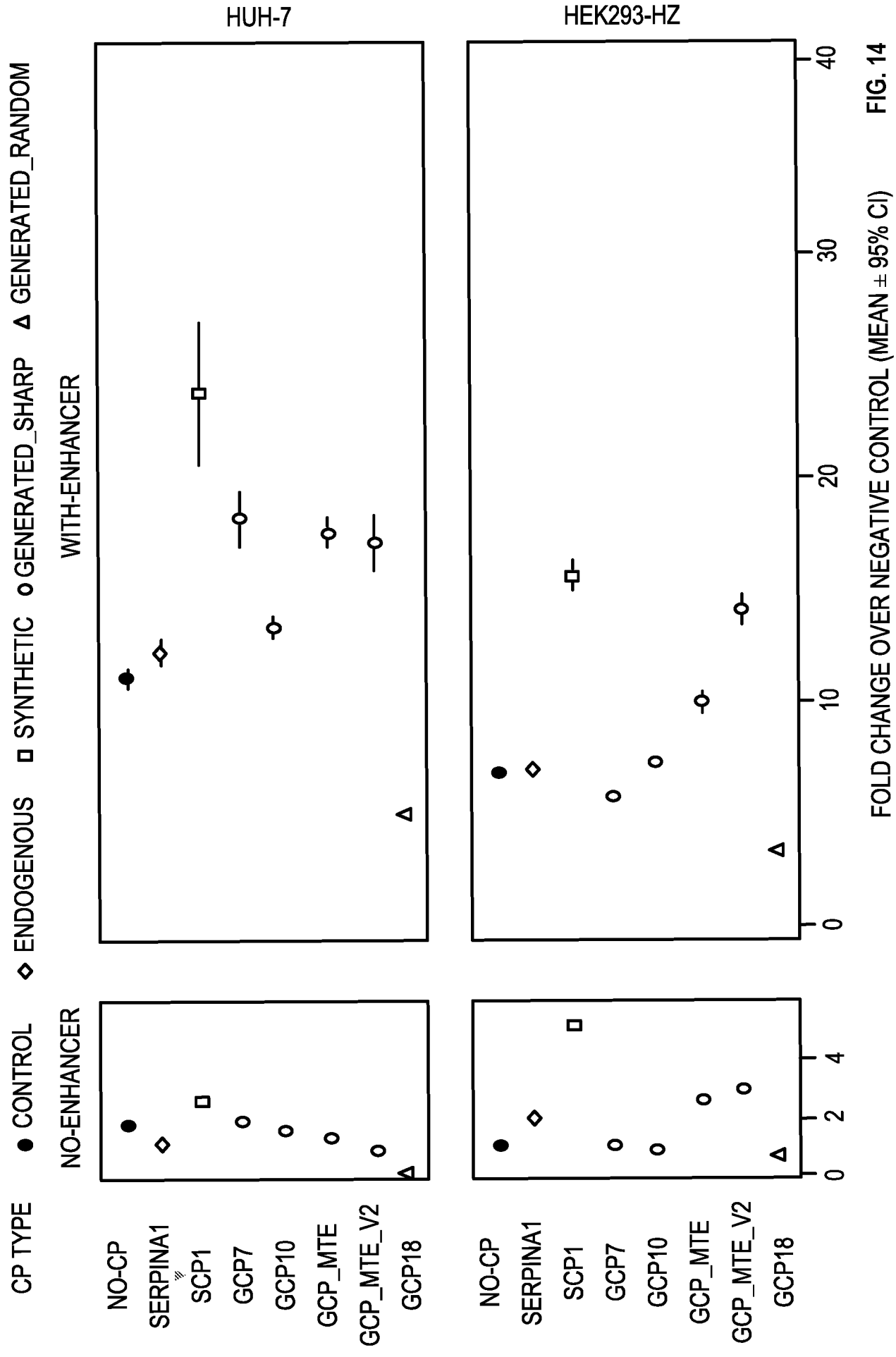

CP TYPE
CONTROL
ENDOGENOUS
SYNTHETIC
GENERATED_SHARP
GENERATED_RANDOM
NO-ENHANCER
WITH-ENHANCER
HUH-7
HEK293-HZ
NO-CP
SERPINA1
SCP1
GCP7
GCP10
GCP_MTE
GCP_MTE_V2
GCP18
CORE PROMOTER
0
2
4
10
20
30
40
FOLD CHANGE OVER NEGATIVE CONTROL (MEAN ± 95% CI)

FIG. 14

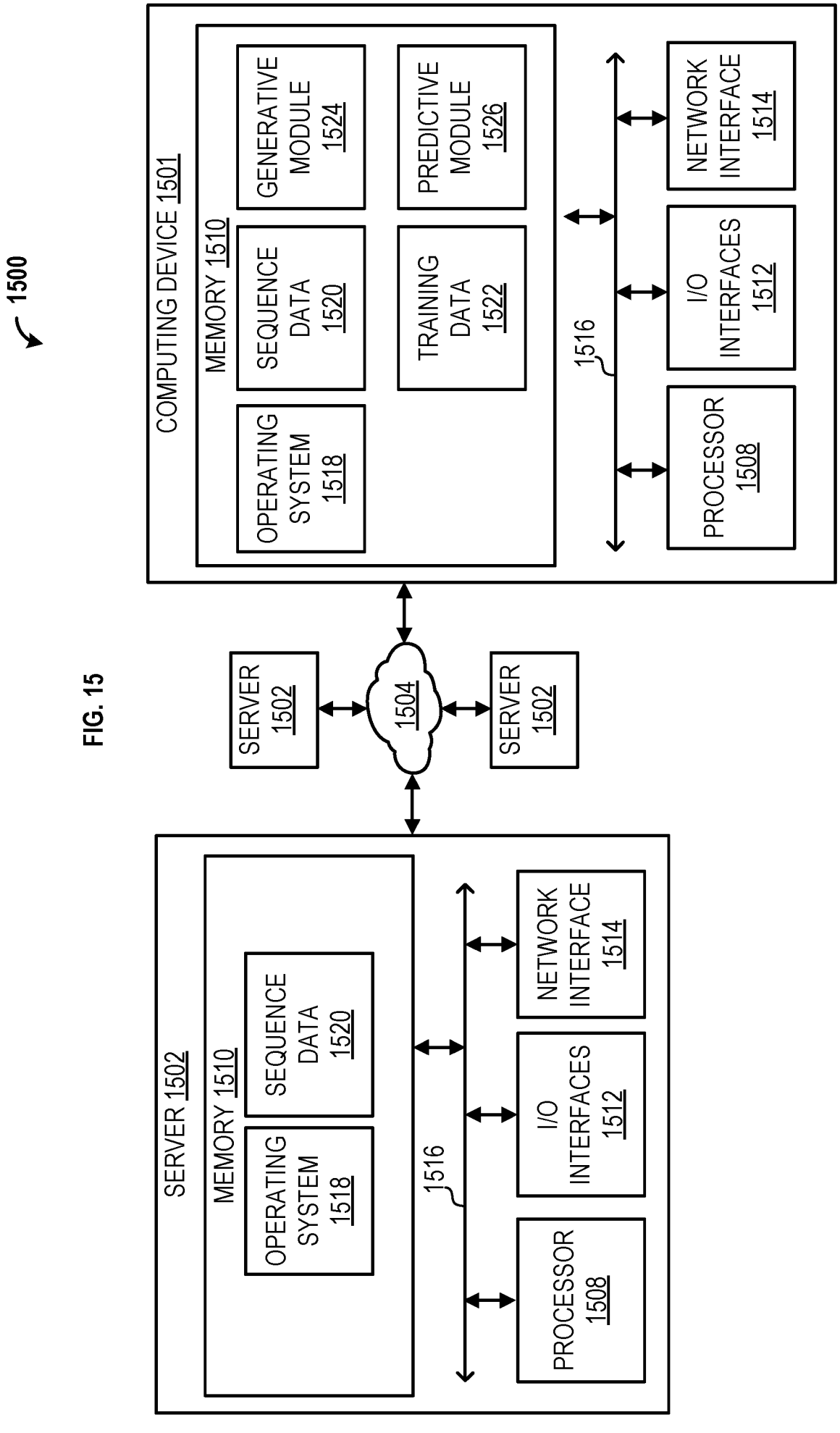
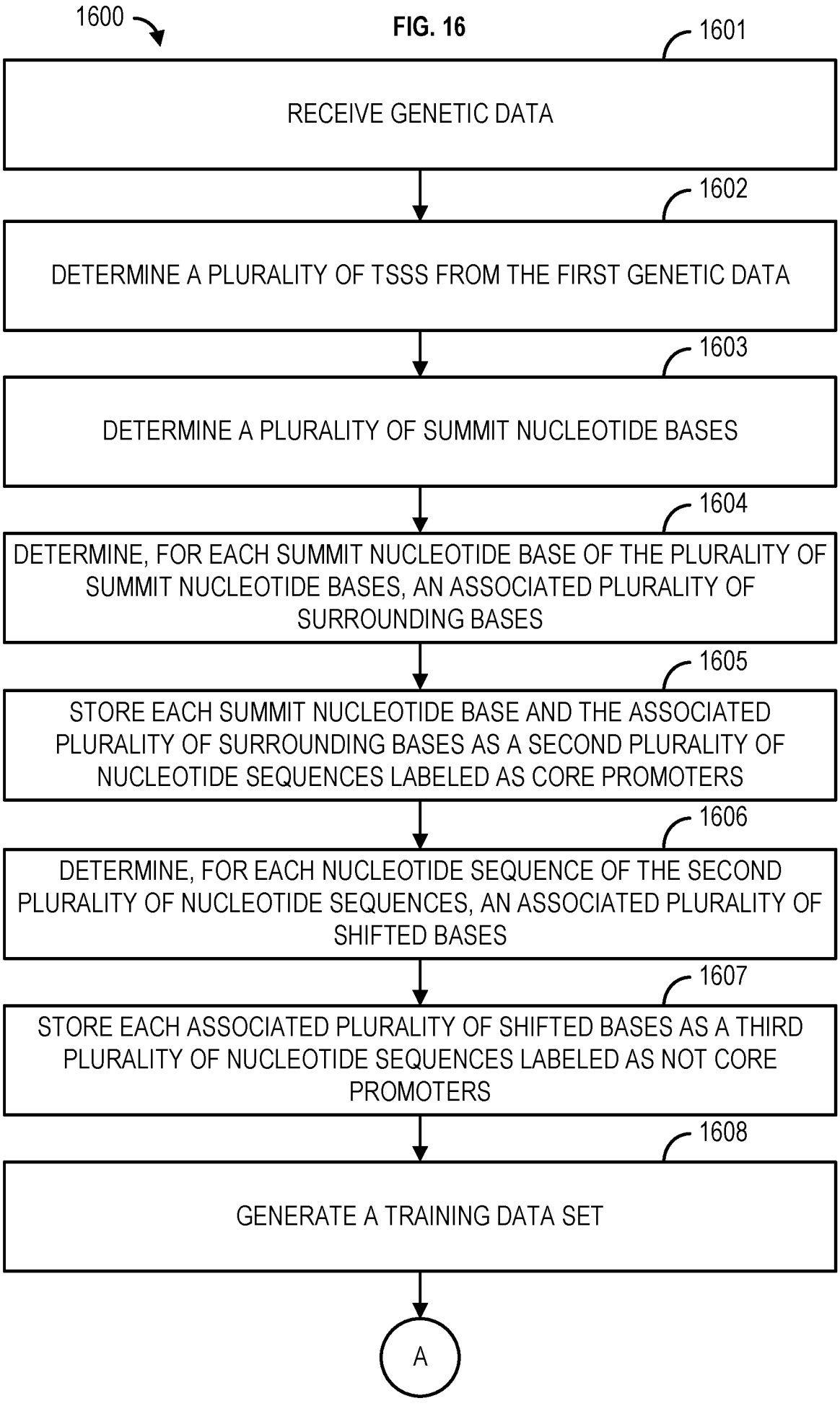

1500
COMPUTING DEVICE 1501
MEMORY 1510
OPERATING SYSTEM 1518
SEQUENCE DATA 1520
TRAINING DATA 1522
GENERATIVE MODULE 1524
PREDICTIVE MODULE 1526
1516
PROCESSOR 1508
I/O INTERFACES 1512
NETWORK INTERFACE 1514
SERVER 1502
1504
SERVER 1502
SERVER 1502
MEMORY 1510
OPERATING SYSTEM 1518
SEQUENCE DATA 1520
1516
PROCESSOR 1508
I/O INTERFACES 1512
NETWORK INTERFACE 1514

FIG. 15

FIG. 16


1600
1601
RECEIVE GENETIC DATA
1602
DETERMINE A PLURALITY OF TSSS FROM THE FIRST GENETIC DATA
1603
DETERMINE A PLURALITY OF SUMMIT NUCLEOTIDE BASES
1604
DETERMINE, FOR EACH SUMMIT NUCLEOTIDE BASE OF THE PLURALITY OF SUMMIT NUCLEOTIDE BASES, AN ASSOCIATED PLURALITY OF SURROUNDING BASES
1605
STORE EACH SUMMIT NUCLEOTIDE BASE AND THE ASSOCIATED PLURALITY OF SURROUNDING BASES AS A SECOND PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS CORE PROMOTERS
1606
DETERMINE, FOR EACH NUCLEOTIDE SEQUENCE OF THE SECOND PLURALITY OF NUCLEOTIDE SEQUENCES, AN ASSOCIATED PLURALITY OF SHIFTED BASES
1607
STORE EACH ASSOCIATED PLURALITY OF SHIFTED BASES AS A THIRD PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS NOT CORE PROMOTERS
1608
GENERATE A TRAINING DATA SET
A

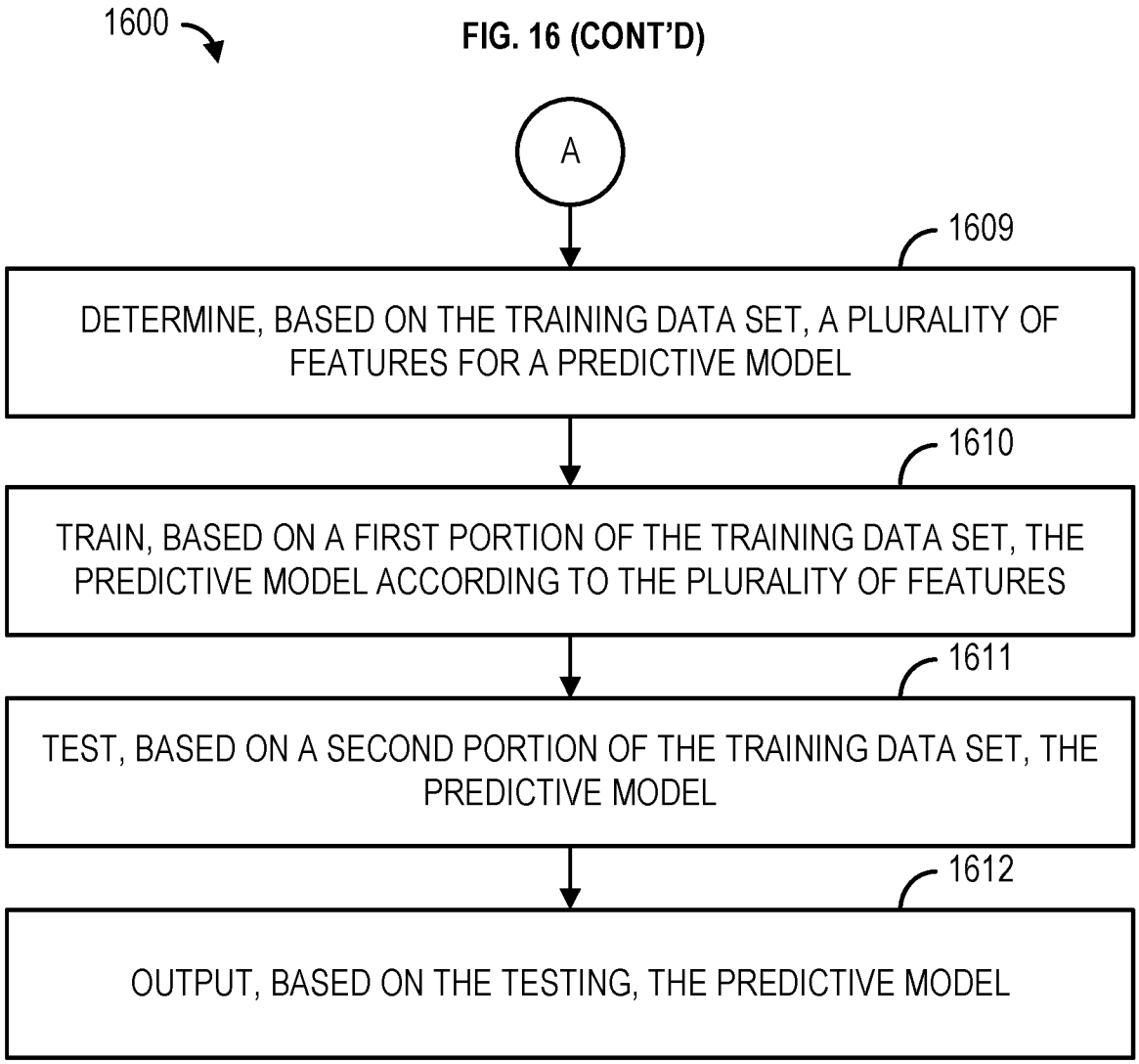

1600
A
1609
DETERMINE, BASED ON THE TRAINING DATA SET, A PLURALITY OF FEATURES FOR A PREDICTIVE MODEL
1610
TRAIN, BASED ON A FIRST PORTION OF THE TRAINING DATA SET, THE PREDICTIVE MODEL ACCORDING TO THE PLURALITY OF FEATURES
1611
TEST, BASED ON A SECOND PORTION OF THE TRAINING DATA SET, THE PREDICTIVE MODEL
1612
OUTPUT, BASED ON THE TESTING, THE PREDICTIVE MODEL

FIG. 16 (CONT'D)

FIG. 17

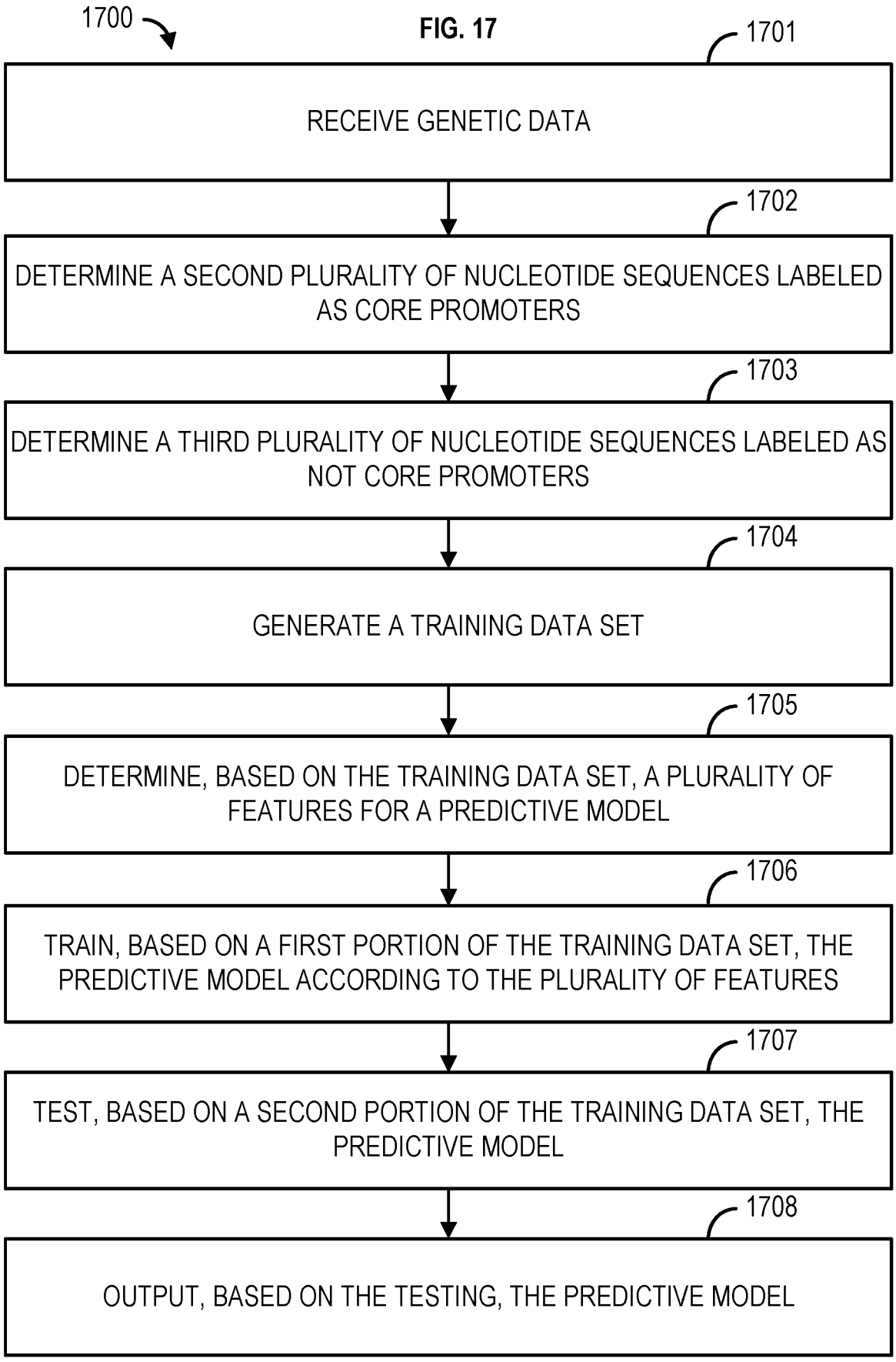
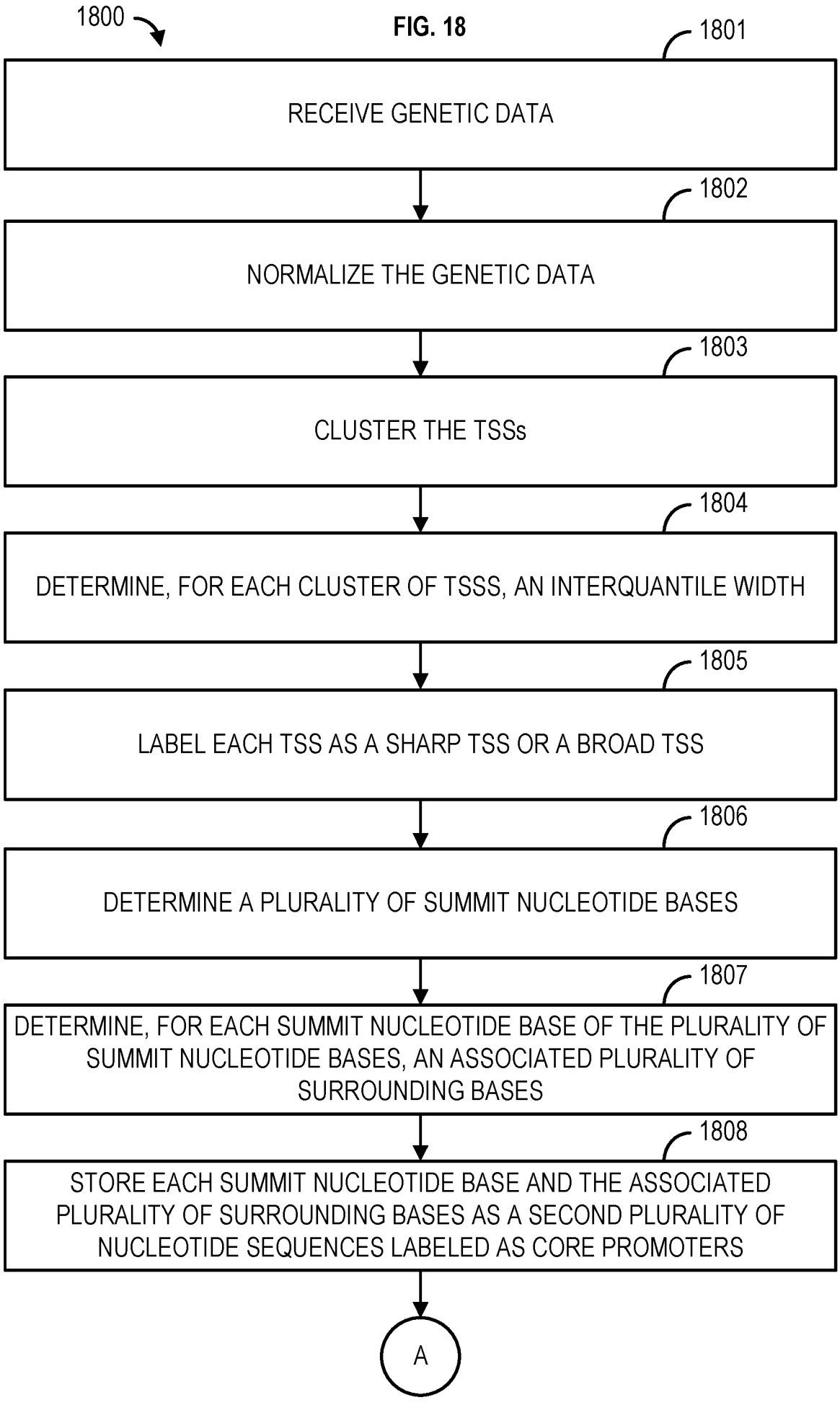

1700
1701
RECEIVE GENETIC DATA
1702
DETERMINE A SECOND PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS CORE PROMOTERS
1703
DETERMINE A THIRD PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS NOT CORE PROMOTERS
1704
GENERATE A TRAINING DATA SET
1705
DETERMINE, BASED ON THE TRAINING DATA SET, A PLURALITY OF FEATURES FOR A PREDICTIVE MODEL
1706
TRAIN, BASED ON A FIRST PORTION OF THE TRAINING DATA SET, THE PREDICTIVE MODEL ACCORDING TO THE PLURALITY OF FEATURES
1707
TEST, BASED ON A SECOND PORTION OF THE TRAINING DATA SET, THE PREDICTIVE MODEL
1708
OUTPUT, BASED ON THE TESTING, THE PREDICTIVE MODEL

FIG. 18


1800
1801
RECEIVE GENETIC DATA
1802
NORMALIZE THE GENETIC DATA
1803
CLUSTER THE TSSs
1804
DETERMINE, FOR EACH CLUSTER OF TSSS, AN INTERQUANTILE WIDTH
1805
LABEL EACH TSS AS A SHARP TSS OR A BROAD TSS
1806
DETERMINE A PLURALITY OF SUMMIT NUCLEOTIDE BASES
1807
DETERMINE, FOR EACH SUMMIT NUCLEOTIDE BASE OF THE PLURALITY OF SUMMIT NUCLEOTIDE BASES, AN ASSOCIATED PLURALITY OF SURROUNDING BASES
1808
STORE EACH SUMMIT NUCLEOTIDE BASE AND THE ASSOCIATED PLURALITY OF SURROUNDING BASES AS A SECOND PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS CORE PROMOTERS
A

1800

FIG. 18 (CONT'D)

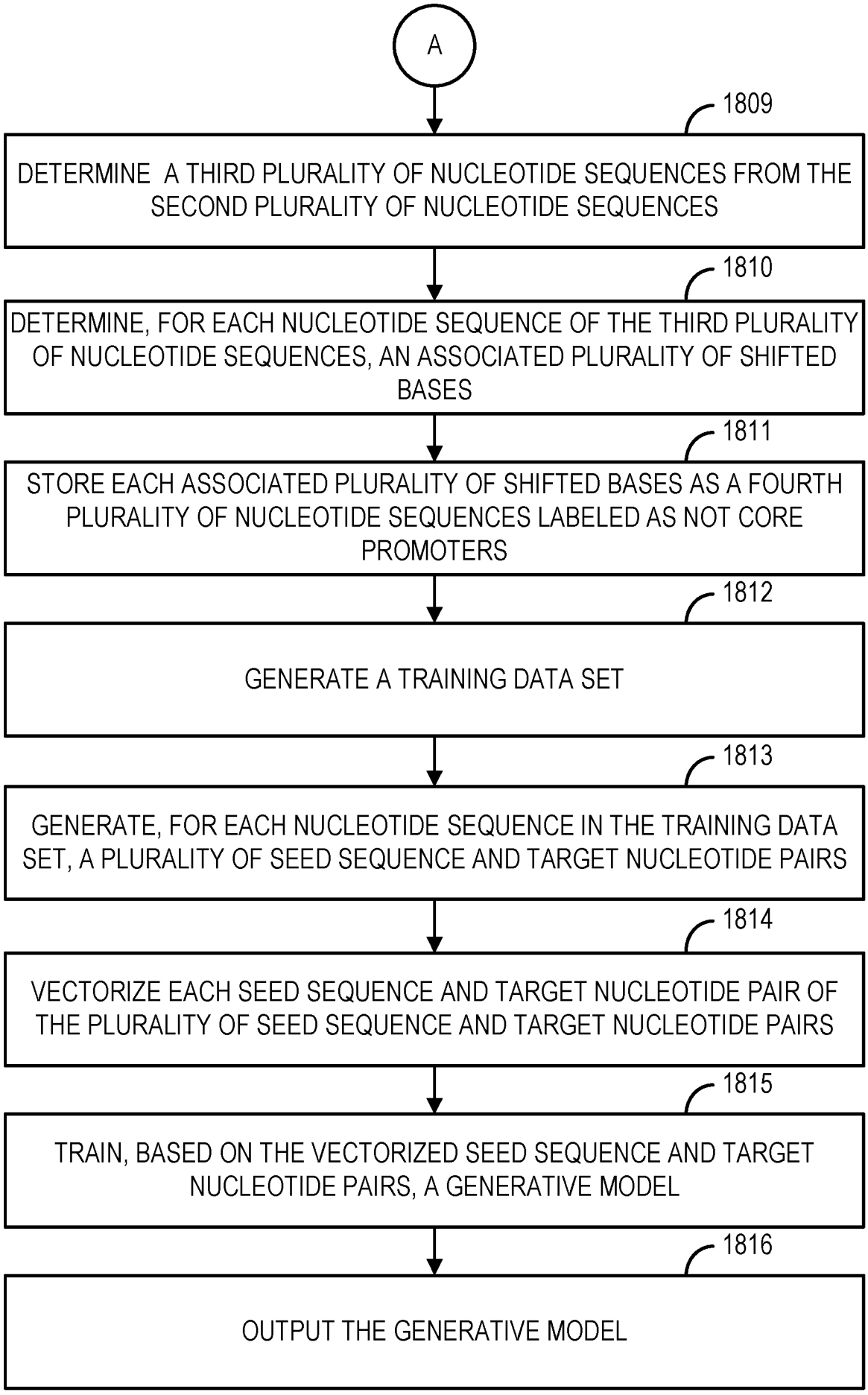
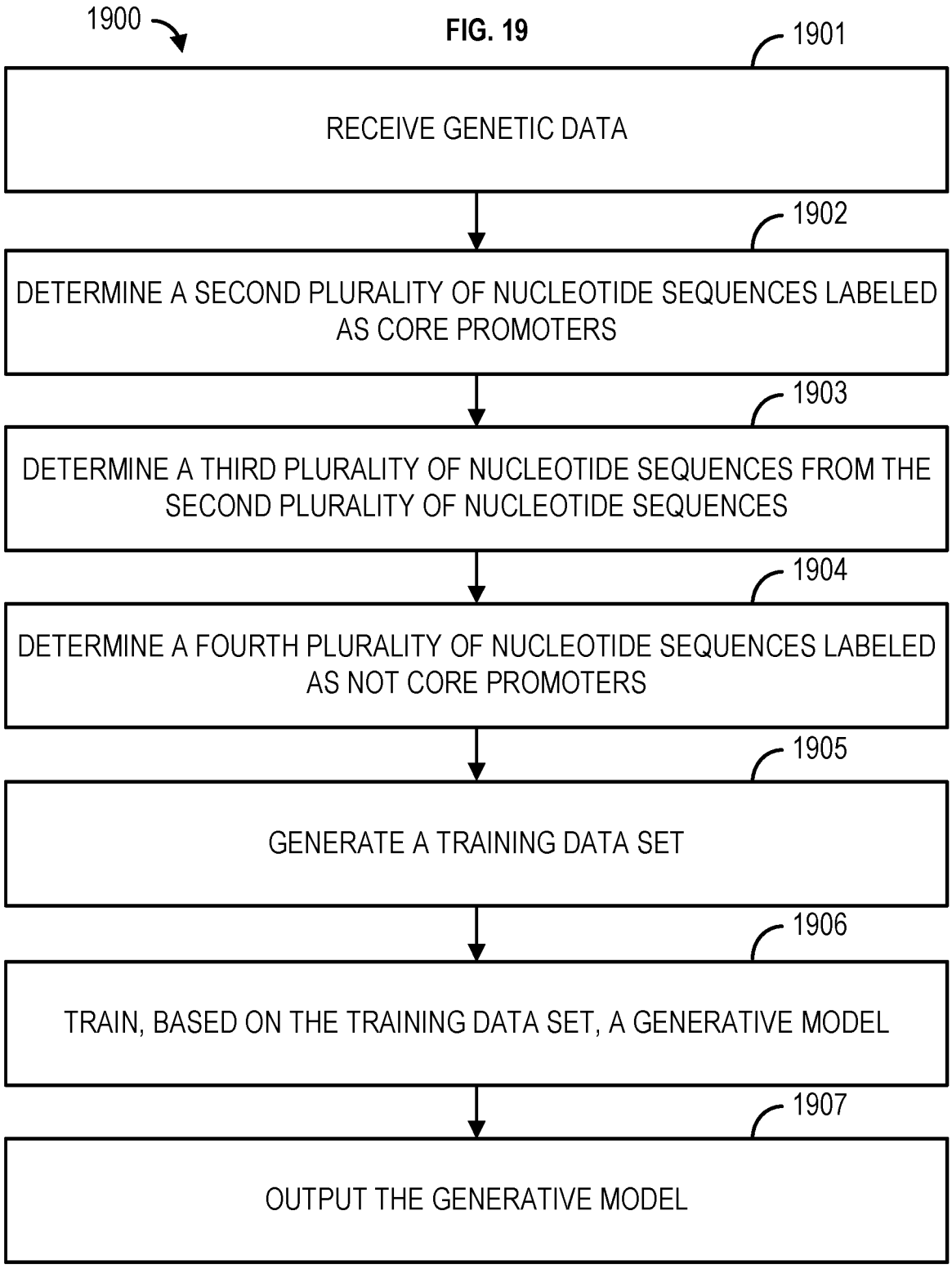


FIG. 19


1900
1901
RECEIVE GENETIC DATA
1902
DETERMINE A SECOND PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS CORE PROMOTERS
1903
DETERMINE A THIRD PLURALITY OF NUCLEOTIDE SEQUENCES FROM THE SECOND PLURALITY OF NUCLEOTIDE SEQUENCES
1904
DETERMINE A FOURTH PLURALITY OF NUCLEOTIDE SEQUENCES LABELED AS NOT CORE PROMOTERS
1905
GENERATE A TRAINING DATA SET
1906
TRAIN, BASED ON THE TRAINING DATA SET, A GENERATIVE MODEL
1907
OUTPUT THE GENERATIVE MODEL

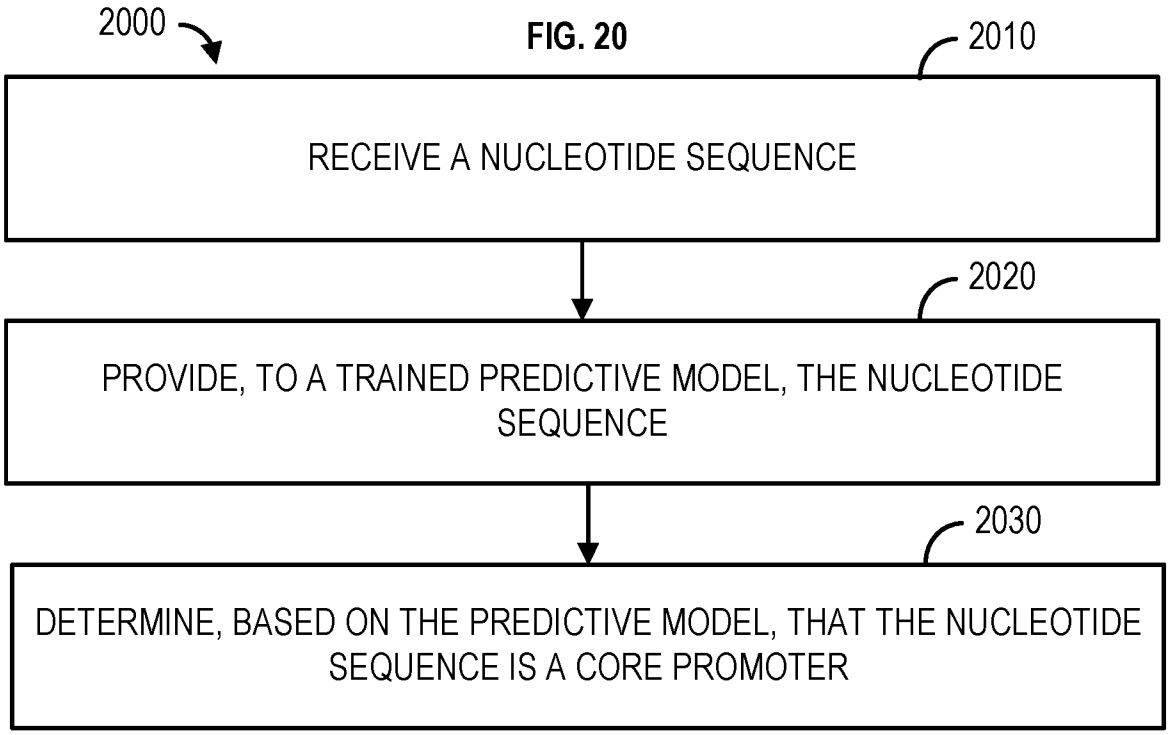

2000
2010
RECEIVE A NUCLEOTIDE SEQUENCE
2020
PROVIDE, TO A TRAINED PREDICTIVE MODEL, THE NUCLEOTIDE SEQUENCE
2030
DETERMINE, BASED ON THE PREDICTIVE MODEL, THAT THE NUCLEOTIDE SEQUENCE IS A CORE PROMOTER

FIG. 20

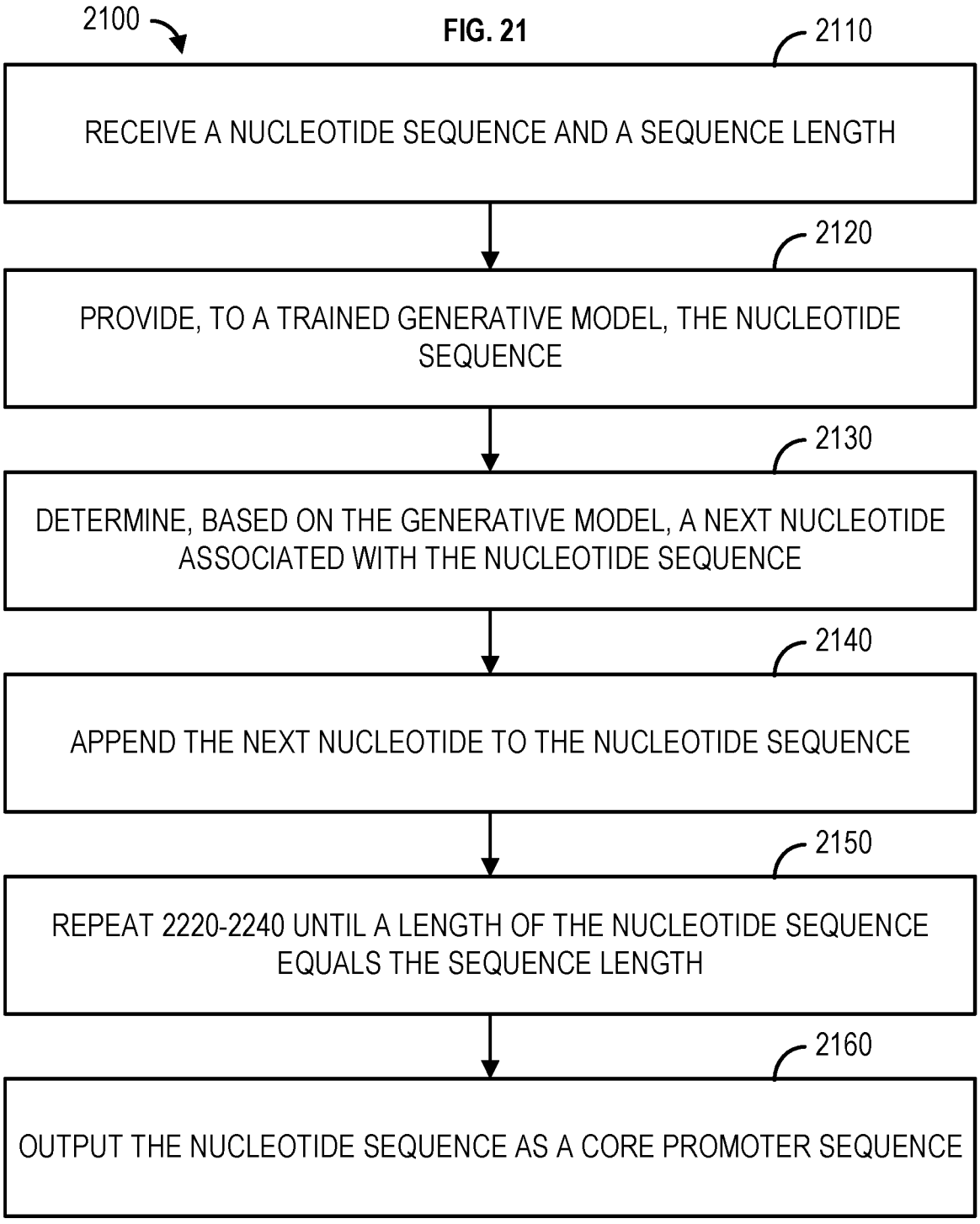

2100
2110
RECEIVE A NUCLEOTIDE SEQUENCE AND A SEQUENCE LENGTH
2120
PROVIDE, TO A TRAINED GENERATIVE MODEL, THE NUCLEOTIDE SEQUENCE
2130
DETERMINE, BASED ON THE GENERATIVE MODEL, A NEXT NUCLEOTIDE ASSOCIATED WITH THE NUCLEOTIDE SEQUENCE
2140
APPEND THE NEXT NUCLEOTIDE TO THE NUCLEOTIDE SEQUENCE
2150
REPEAT 2220-2240 UNTIL A LENGTH OF THE NUCLEOTIDE SEQUENCE EQUALS THE SEQUENCE LENGTH
2160
OUTPUT THE NUCLEOTIDE SEQUENCE AS A CORE PROMOTER SEQUENCE

FIG. 21