ARCHITECTURE-ALTERING OPERATIONS

FINDING THE ARCHITECTURE OF THE AUTOMATICALLY DEFINED FUNCTIONS

MANUAL METHODS

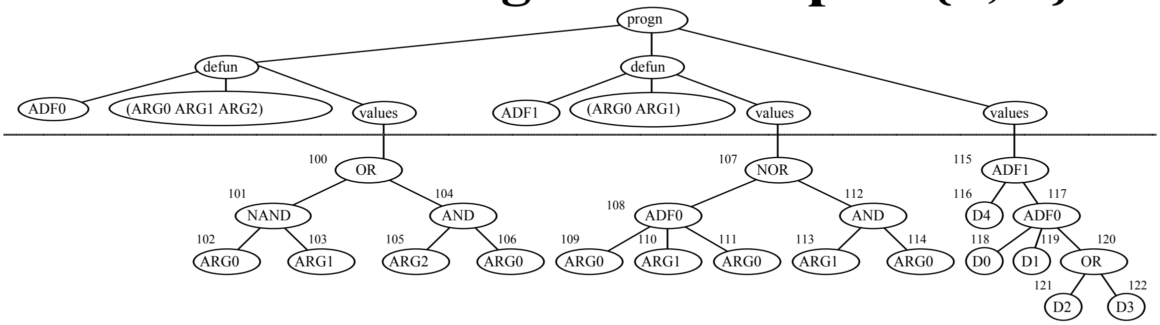
- **prospective analysis of the problem**
- **seemingly sufficient capacity (over-specification)**
- **affordable capacity**
- **retrospective analysis of the results of actual runs**

AUTOMATED METHODS

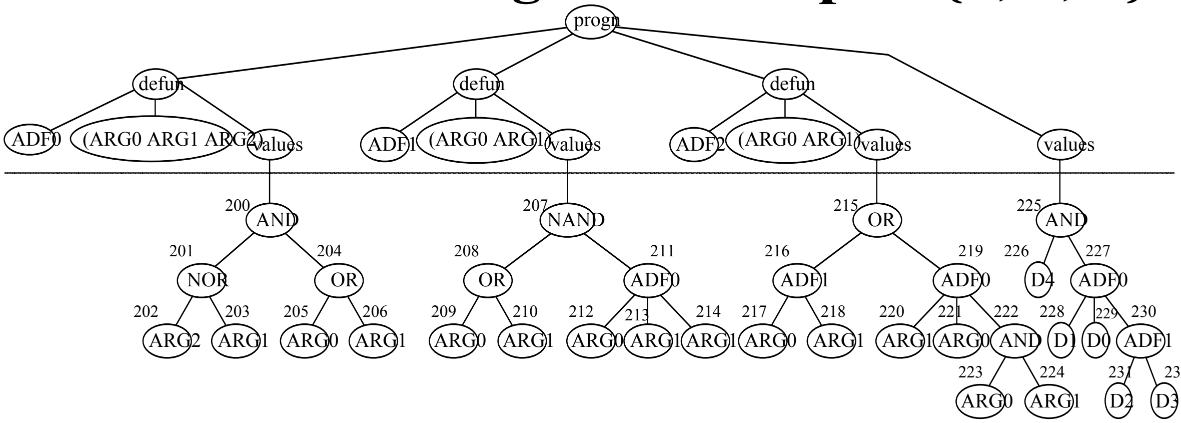
- **evolutionary selection of the architecture**
- **evolution of architecture using architecture-altering operations**

ARCHITECTURALLY DIVERSE POPULATION

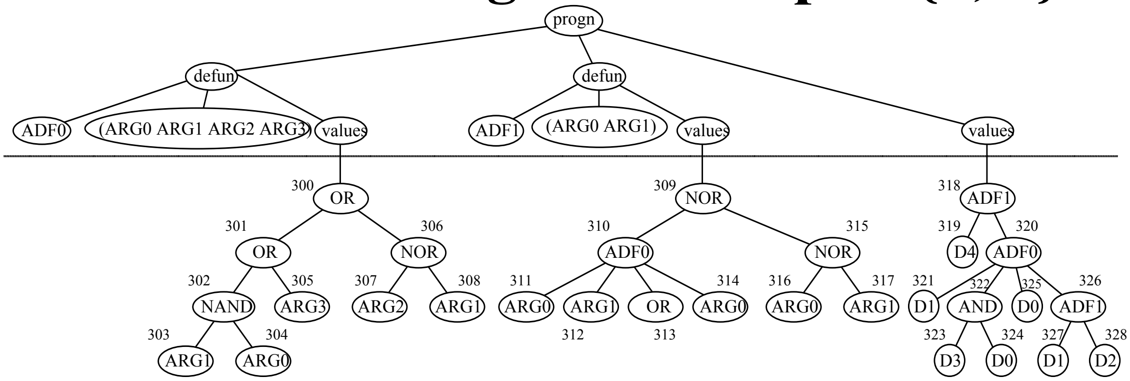
Parent A with an argument map of {3, 2}



Parent B with an argument map of {3, 2, 2}



Parent C with an argument map of {4, 2}



POINT TYPING – STRUCTURE-PRESERVING CROSSOVER

GENE DUPLICATION IN NATURE

- **Midge *Chironomus tentans* (Galli and Wislander 1993)**
- **3,959-bases of DNA with accession number X70063 in GenBank**
- **One subsequence of 732 bases (called "C. tentans Sp38–40.A gene") are in DNA positions positions 918–1,649 and is expressed as protein of length 244**
- **A second subsequence of 759 bases (called "C. tentans Sp38–40.B gene") are in DNA positions 2,513–3,271 and is expressed as protein of length 253.**
- **Both proteins are secreted from the salivary gland of the insect and form water-insoluble fibers which are spun into one of two kinds of tubes – one for larval protection and feeding and one for pupation**

MIDGE *CHIRONOMUS TENTANS*

```
TGAAGTAATA   TTAAGCTATG   AGAATTAAGT   TCCTAGTAGT   ATTAGCAGTT   950
                    M     R  I  K  F     L  V  V     L  A  V

ATCTGCTTGT   TTGCACATTA   TGCCTCAGCT   AGTGGTATGG   GGGGTGATAA   1000
I  C  L  F     A  H  Y     A  S  A     S  G  M  G     G  D  K

AAAACCCAAA   GATGCCCCAA   AACCCAAAGA   TGCCCCAAAA   CCCAAAGAAG   1050
 K  P  K     D  A  P  K     P  K  D     A  P  K     P  K  E  V

TGAAGCCTGT   CAAAGCTGAG   TCATCAGAGT   ATGAGATAGA   AGTCATTAAA   1100
  K  P  V     K  A  E     S  S  E  Y     E  I  E     V  I  K

CACCAGAAAG   AAAAGACCGA   GAAGAAGGAG   AAGGAGAAGA   AGACTCACGT   1150
H  Q  K  E     K  T  E     K  K  E     K  E  K  K     T  H  V

TGAAACCAAG   AAAGAAGTTA   AAAAGAAGGA   GAAGAAGCAA   ATCCCTTGTT   1200
 E  T  K     K  E  V  K     K  K  E     K  K  Q     I  P  C  S

CTGAAAAACT   CAAGGATGAA   AAACTTGATT   GTGAGACCAA   GGGCGTCCCT   1250
  E  K  L     K  D  E     K  L  D  C     E  T  K     G  V  P

GCAGGCTACA   AAGCAATCTT   CAAATTCACA   GAAAACGAGG   AGTGCGATTG   1300
A  G  Y  K     A  I  F     K  F  T     E  N  E  E     C  D  W

GACGTGCGAT   TATGAAGCAC   TTCCACCACC   TCCAGGAGCA   AAGAAAGACG   1350
 T  C  D     Y  E  A  L     P  P  P     P  G  A     K  K  D  D

ACAAGAAAGA   AAAGAAGACA   GTTAAAGTCG   TTAAGCCACC   AAAGGAGAAA   1400
  K  K  E     K  K  T     V  K  V  V     K  P  P     K  E  K

CCACCAAAGA   AGCTTAGAAA   GGAATGCTCT   GGCGAAAAAG   TGATCAAATT   1450
P  P  K  K     L  R  K     E  C  S     G  E  K  V     I  K  F

CCAAAACTGT   CTCGTTAAGA   TTAGAGGACT   TATTGCCTTT   GGTGATAAGA   1500
 Q  N  C     L  V  K  I     R  G  L     I  A  F     G  D  K  T

CAAAGAACTT   TGATAAGAAG   TTCGCAAAGC   TTGTCCAAGG   AAAGCAGAAG   1550
  K  N  F     D  K  K     F  A  K  L     V  Q  G     K  Q  K

AAGGGCGCAA   AAAAAGCTAA   AGGCGGTAAG   AAGGCAGCAC   CAAAACCAGG   1600
K  G  A  K     K  A  K     G  G  K     K  A  A  P     K  P  G

ACCAAAACCA   GGGCCAAAAC   AAGCTGATAA   ACCAAAAGAT   GCAAAAAAAT   1650
 P  K  P     G  P  K  Q     A  D  K     P  K  D     A  K  K

AAACTGACAT   AGTAAGAATA   ATAAAATAAA   CATTATTTGA   GCAACATCAC   1700
AACACAAGAA   AAAAATCATA   TCAACATAAT   TAAGACCTAA   AAATTCTCGC   1750
TATTCACTTT   TTTTCAAATG   AATATCCAAA   ACAACATCAT   TAAGGGATCT   1800
TACACAATTT   TATCCCAAAT   TAGTTTTAAG   TCTATTTTTT   AGTTTTAAGT   1850
AAAACATTAG   TTAGAGAAAT   TTCAAATGCG   AAAAAAAGAC   AAAATCAAAA   1900
TTAACTCCAA   CTAATTGTCT   AGATCTAATC   ACCACTGAAA   AACAATATTT   1950
TTTTCAATAA   TATCTGAGAT   GAAAATTTTG   TAAGATACGA   TTCAAAAAAA   2000
AAAAAACAAA   AACTTAAATA   TTTTCTTTAT   AAGAAAGTAA   AAAACTTACA   2050
TGAACAACAA   GTAGACTAAG   GGCTTAAAAA   TACTAAGGAA   TTTAAAGAAA   2100
CTGAACCAAT   AACATCCAAT   AAATATAAGC   GTGTATTTAA   CATCCATTCA   2150
TGCAAAATTT   GACTTGTTTT   ATTCTAAACT   TTTGAATTGT   GAATATTTTT   2200
GATGATTATT   GAATATTTTA   CAGCATTTTT   CGACAAAATC   CAAGGAAACT   2250
GTTTTGTTTA   ATATATACTA   CAGCTCAGTA   TCTATGCACA   CGAAAAACTG   2300
TAACAGACCA   GACCATAAAA   CCTACACATC   ACCAAGATAC   GTATTTTAAA   2350
TTCATGTGAC   TGACAAAAGC   TGGAAACACT   TGTGTCACGT   CATGAAAACC   2400
TCGTTGAAAT   AAAACTTCTA   GAAAGGTTAT   CATGAAAGAG   TATAAAAGAG   2450
ATCTCAAACG   AGGCTCAGTC   AGTTCAGTTT   AGCTTGGACT   TCATATGAAG   2500
```

```
TAATATTTAG   CTATGAGAAT   TAAGTTCCTA   GTAGTATTAG   CAGTTATCTG   2550
               M  R  I     K  F  L     V  V  L  A     V  I  C

CTTGCTTGCA   CATTATGCCT   CAGCTAGTGG   TATGGGGGGT   GATAAAAAAC   2600
 L  L  A     H  Y  A  S     A  S  G     M  G  G     D  K  K  P

CCAAAGATGC   CCCAAAACCC   AAAGATGCCC   CAAAACCCAA   AGAAGTGAAG   2650
  K  D  A     P  K  P     K  D  A  P     K  P  K     E  V  K

CCTGTCAAAG   CTGACTCATC   AGAGTATGAG   ATAGAAGTCA   TTAAACACCA   2700
P  V  K  A     D  S  S     E  Y  E     I  E  V  I     K  H  Q

GAAAGAAAAG   ACCGAGAAGA   AGGAGAAGGA   GAAGAAAGCT   CACGTCGAAA   2750
 K  E  K     T  E  K  K     E  K  E     K  K  A     H  V  E  I

TCAAGAAAAA   GATTAAAAAT   AAGGAGAAGA   AGTTTGTCCC   ATGTTCTGAA   2800
  K  K  K     I  K  N     K  E  K  K     F  V  P     C  S  E

ATTCTCAAGG   ATGAAAAACT   TGAATGTGAG   AAAAATGCTA   CTCCAGGCTA   2850
I  L  K  D     E  K  L     E  C  E     K  N  A  T     P  G  Y

TAAAGCACTC   TTCGAATTCA   AAGAAAGCGA   AAGTTTTTGC   GAATGGGAGT   2900
 K  A  L     F  E  F  K     E  S  E     S  F  C     E  W  E  C

GCGATTATGA   AGCAATTCCA   GGAGCAAAGA   AAGACGAAAA   AAAGGAGAAG   2950
  D  Y  E     A  I  P     G  A  K  K     D  E  K     K  E  K

AAGGTAGTTA   AAGTCATTAA   GCCACCAAAG   GAAAAACCAC   CAAAGAAGCC   3000
K  V  V  K     V  I  K     P  P  K     E  K  P  P     K  K  P

TAGAAAGGAA   TGCTCTGGCG   AAAAAGTGAT   CAAATTCCAA   AACTGTCTCG   3050
 R  K  E     C  S  G  E     K  V  I     K  F  Q     N  C  L  V

TTAAGATTAG   AGGACTTATT   GCCTTTGGTG   ATAAGACAAA   GAACTTTGAT   3100
  K  I  R     G  L  I     A  F  G  D     K  T  K     N  F  D

AAGAAGTTTG   CAAAGCTTGT   CCAAGGAAAG   CAAAAGAAGG   GCGCAAAAAA   3150
K  K  F  A     K  L  V     Q  G  K     Q  K  K  G     A  K  K

AGCTAAAGGC   GGTAAGAAGG   CAGAACCAAA   ACCAGGACCA   AAACCAGCAC   3200
 A  K  G     G  K  K  A     E  P  K     P  G  P     K  P  A  P

CAAAACCAGG   ACCAAAACCA   GCACCAAAAC   CAGTACCAAA   ACCAGCTGAT   3250
  K  P  G     P  K  P     A  P  K  P     V  P  K     P  A  D

AAACCAAAAG   ATGCAAAAAA   ATAAACTGAC   ATAGTGAGAA   TAATAAAATA   3300
K  P  K  D     A  K  K
```

PROTEIN SEQUENCE OF "A" PROTEIN

```
MRIKFLVVLA   VICLFAHYAS   ASGMGGDKKP   KDAPKPKDAP   KPKEVKPVKA   50
ESSEYEIEVI   KHQKEKTEKK   EKEKKTHVET   KKEVKKKEKK   QIPCSEKLKD   100
EKLDCETKGV   PAGYKAIFKF   TENEECDWTC   DYEALPPPPG   AKKDDKKEKK   150
TVKVVKPPKE   KPPKKLRKEC   SGEKVIKFQN   CLVKIRGLIA   FGDKTKNFDK   200
KFAKLVQGKQ   KKGAKKAKGG   KKAAPKPGPK   PGPKQADKPK   DAKK         244
```

PROTEIN SEQUENCE OF "B" PROTEIN

```
MRIKFLVVLA   VICLLAHYAS   ASGMGGDKKP   KDAPKPKDAP   KPKEVKPVKA   50
DSSEYEIEVI   KHQKEKTEKK   EKEKKAHVEI   KKKIKNKEKK   FVPCSEILKD   100
EKLECEKNAT   PGYKALFEFK   ESESFCEWEC   DYEAIPGAKK   DEKKEKKVVK   150
VIKPPKEKPP   KKPRKECSGE   KVIKFQNCLV   KIRGLIAFGD   KTKNFDKKFA   200
KLVQGKQKKG   AKKAKGGKKA   EPKPGPKPAP   KPGPKPAPKP   VPKPADKPKD   250
AKK                                                            253
```

PROTEIN ALIGNMENT OF "A" AND "B" PROTEINS

```
First.protein     MRIKFLVVLA VICLFAHYAS ASGMGGDKKP KDAPKPKDAP KPKEVKPVKA    50
Second.protein    MRIKFLVVLA VICLLAHYAS ASGMGGDKKP KDAPKPKDAP KPKEVKPVKA    50

First.protein     ESSEYEIEVI KHQKEKTEKK EKEKKTHVET KKEVKKKEKK QIPCSEKLKD   100
Second.protein    DSSEYEIEVI KHQKEKTEKK EKEKKAHVEI KKKIKNKEKK FVPCSEILKD   100

First.protein     EKLDCETKGV PAGYKAIFKF TENEE-CDWT CDYEALPPPP GAKKDDKKEK   149
Second.protein    EKLECEKNAT P-GYKALFEF KESESFCEWE CDYEAI---P GAKKDEKKEK   146

First.protein     KTVKVMKPPK EKPPKKLRKE CSGEKVIKFQ NCLVKIRGLI AFGDKTKNFD   199
Second.protein    KVVKVIKPPK EKPPKKPRKE CSGEKVIKFQ NCLVKIRGLI AFGDKTKNFD   196

First.protein     KKFAKLVQGK QKKGAKKAKG GKKAAPKPGP KPGPK----Q ADKP------   239
Second.protein    KKFAKLVQGK QKKGAKKAKG GKKAEPKPGP KPAPKPGPKP APKPVPKPAD   246

First.protein     --KDAKK                                                  244
Second.protein    KPKDAKK                                                  253
```

NEW ARCHITECTURE-ALTERING OPERATORS

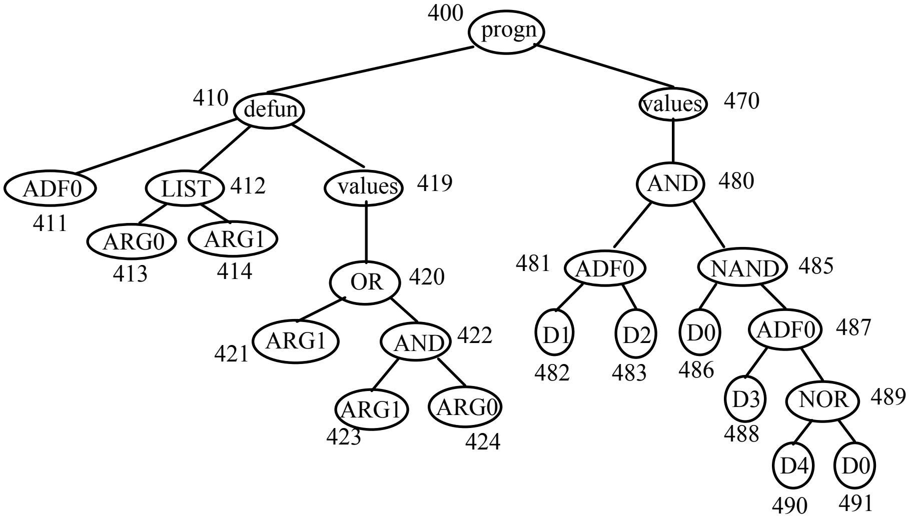
SPECIALIZATION – REFINEMENT – CASE SPLITTING

- **Subroutine (branch) duplication**
- **Argument duplication**
- **Subroutine (branch) creation**
- **Argument creation**

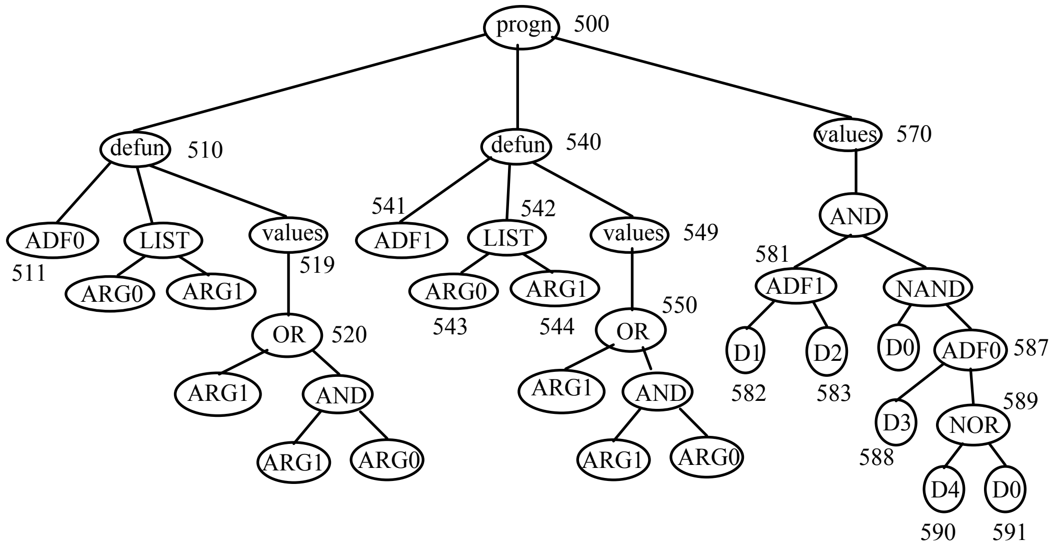
GENERALIZATION

- **Subroutine (branch) deletion**
- **Argument deletion**

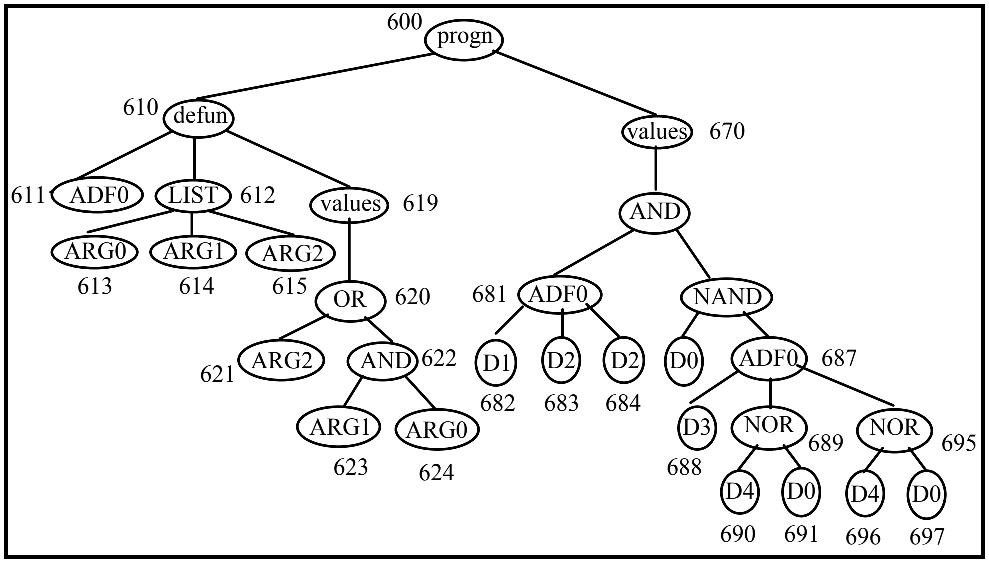
PROGRAM WITH 1 TWO-ARGUMENT AUTOMATICALLY DEFINED FUNCTION (ADF0) AND 1 RESULT-PRODUCING BRANCH – ARGUMENT MAP OF {2}



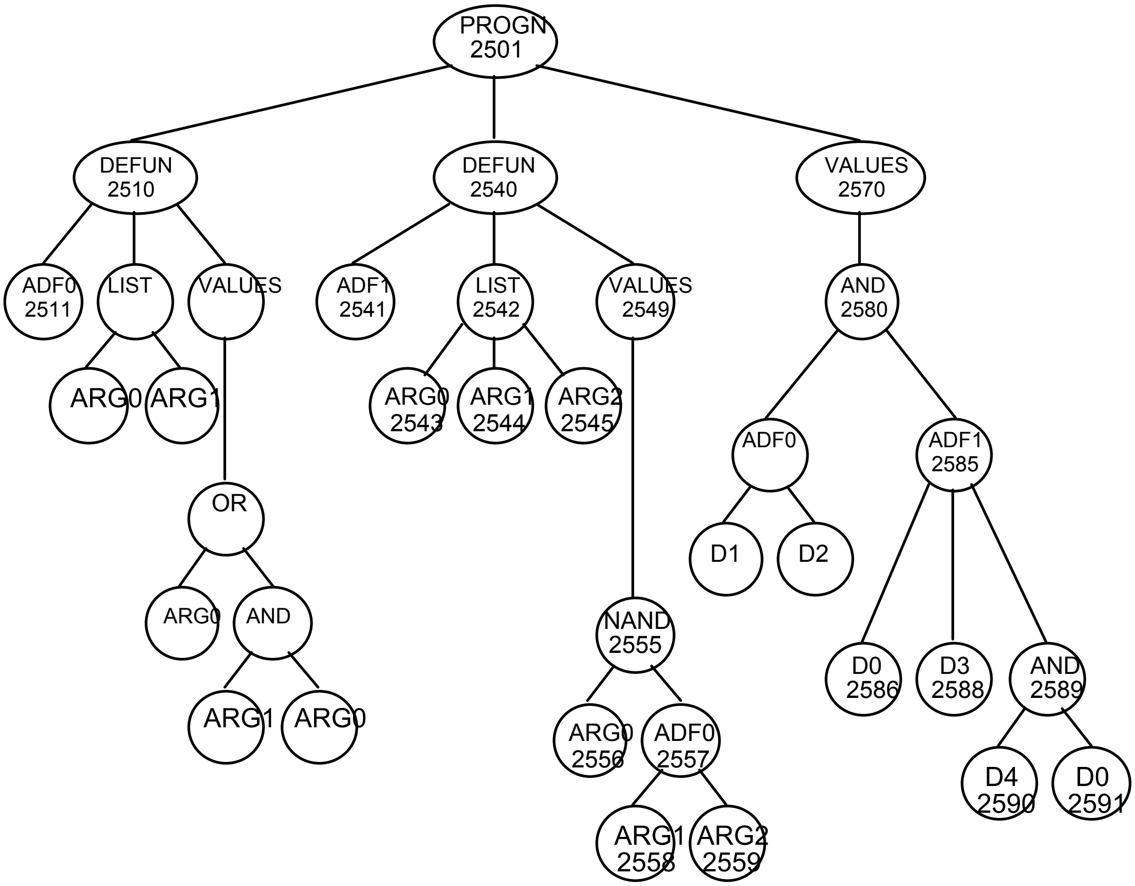
PROGRAM WITH ARGUMENT MAP OF {2, 2} CREATED USING THE OPERATION OF BRANCH DUPLICATION



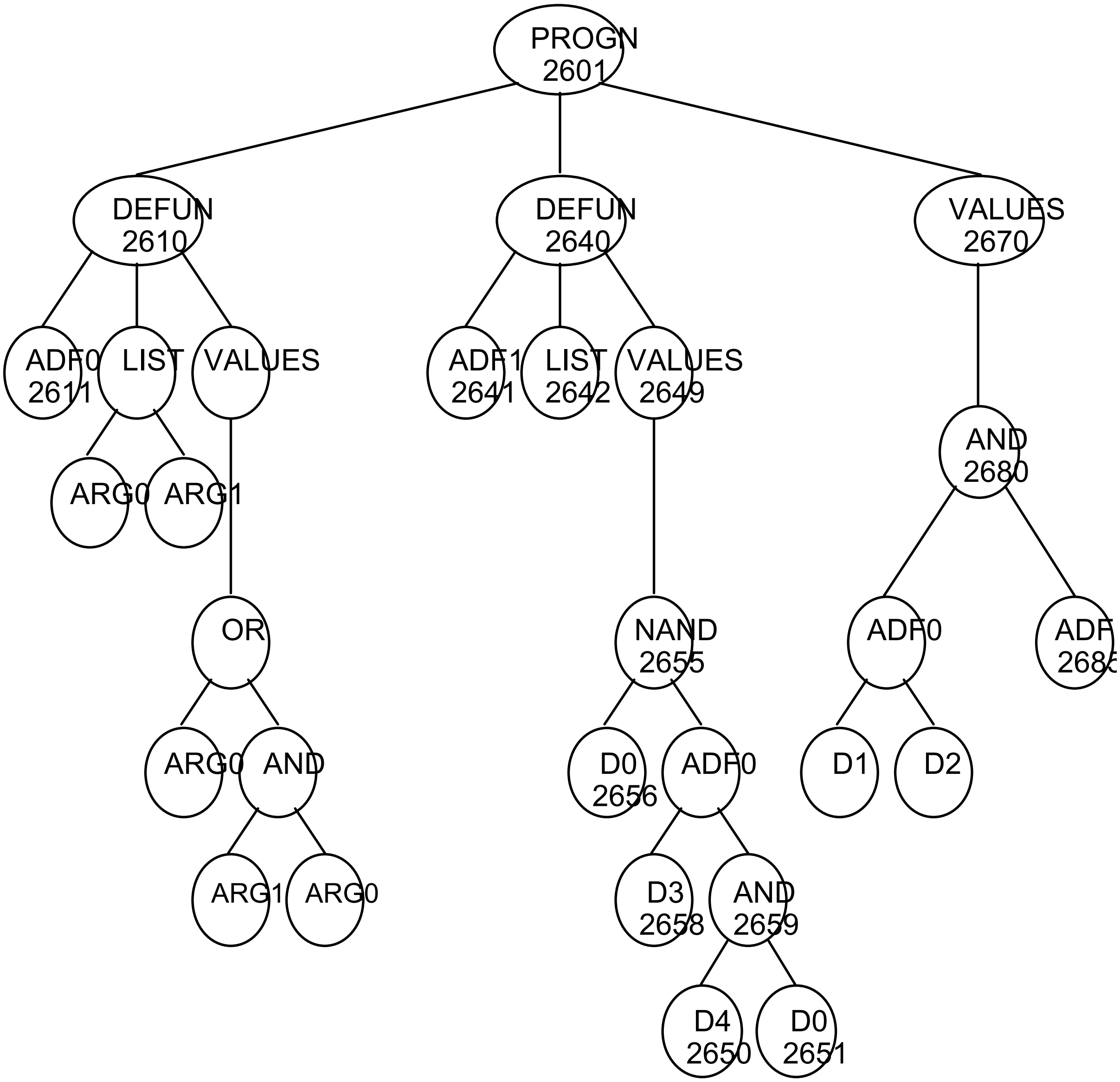
PROGRAM WITH ARGUMENT MAP OF {3) CREATED USING THE OPERATION OF ARGUMENT DUPLICATION



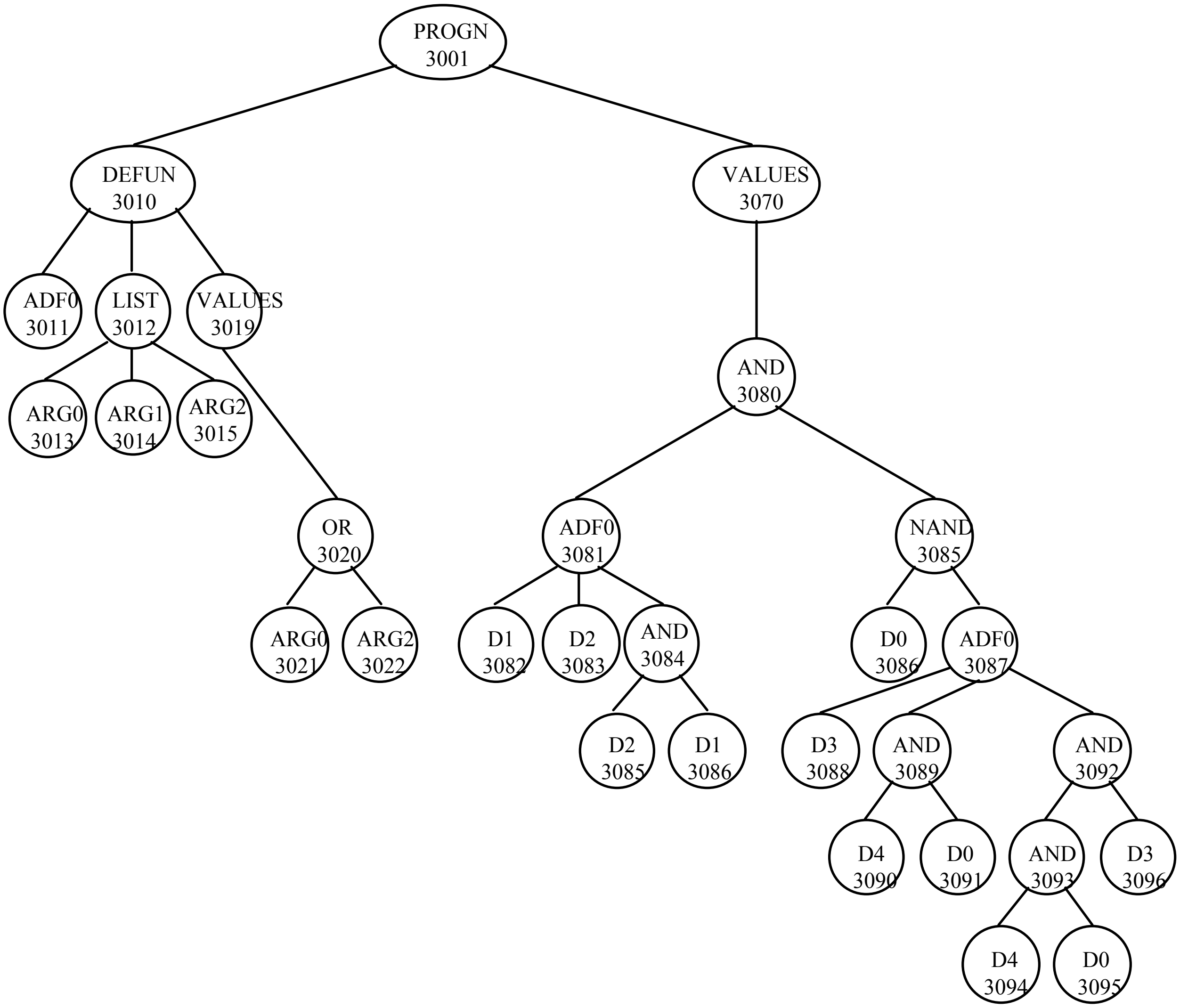
PROGRAM WITH ARGUMENT MAP OF {2, 3}



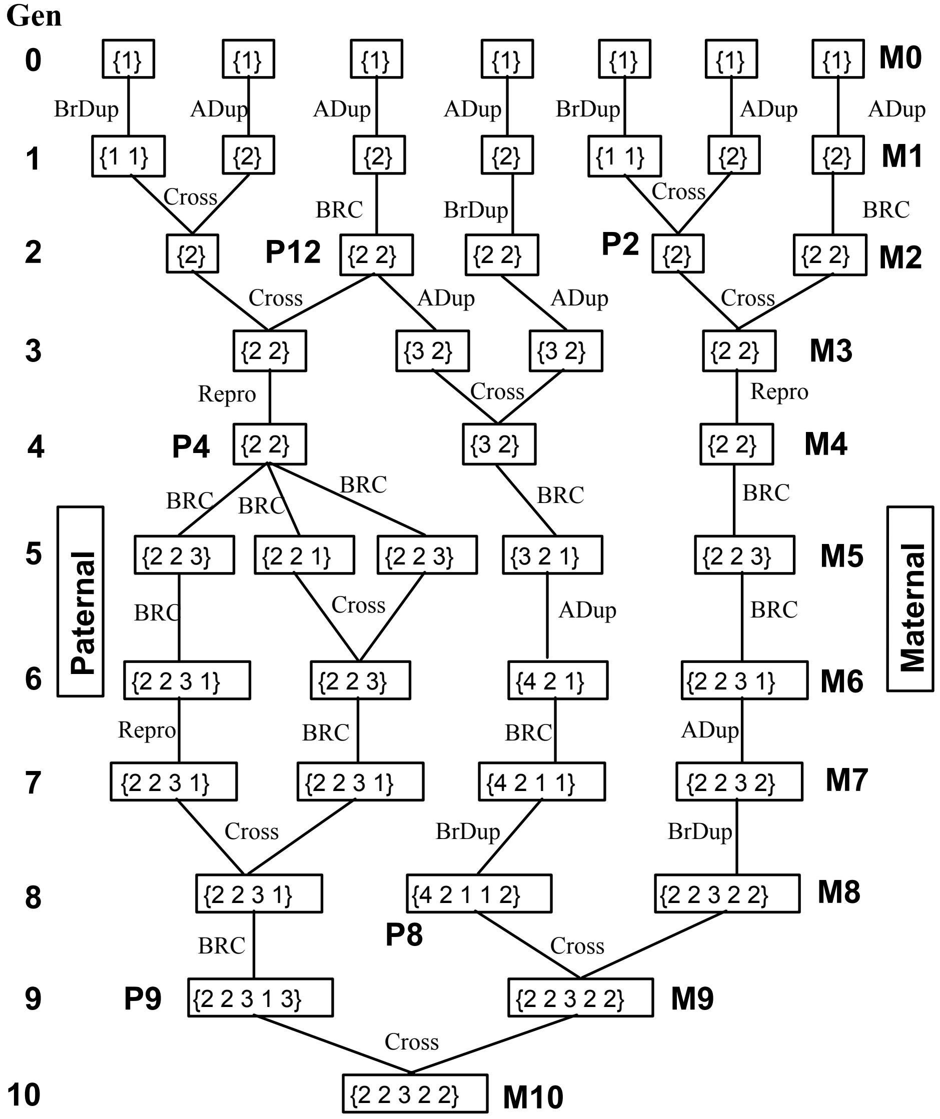
PROGRAM WITH ARGUMENT MAP OF {2, 0}



{3} PROGRAM – ARGUMENT CREATION



AUDIT TRAIL FOR 5-PARITY SOLUTION



EVEN-3-PARITY PROBLEM – BEST-OF-GENERATION 0 - RAW FITNESS OF 6 (OUT OF 8)

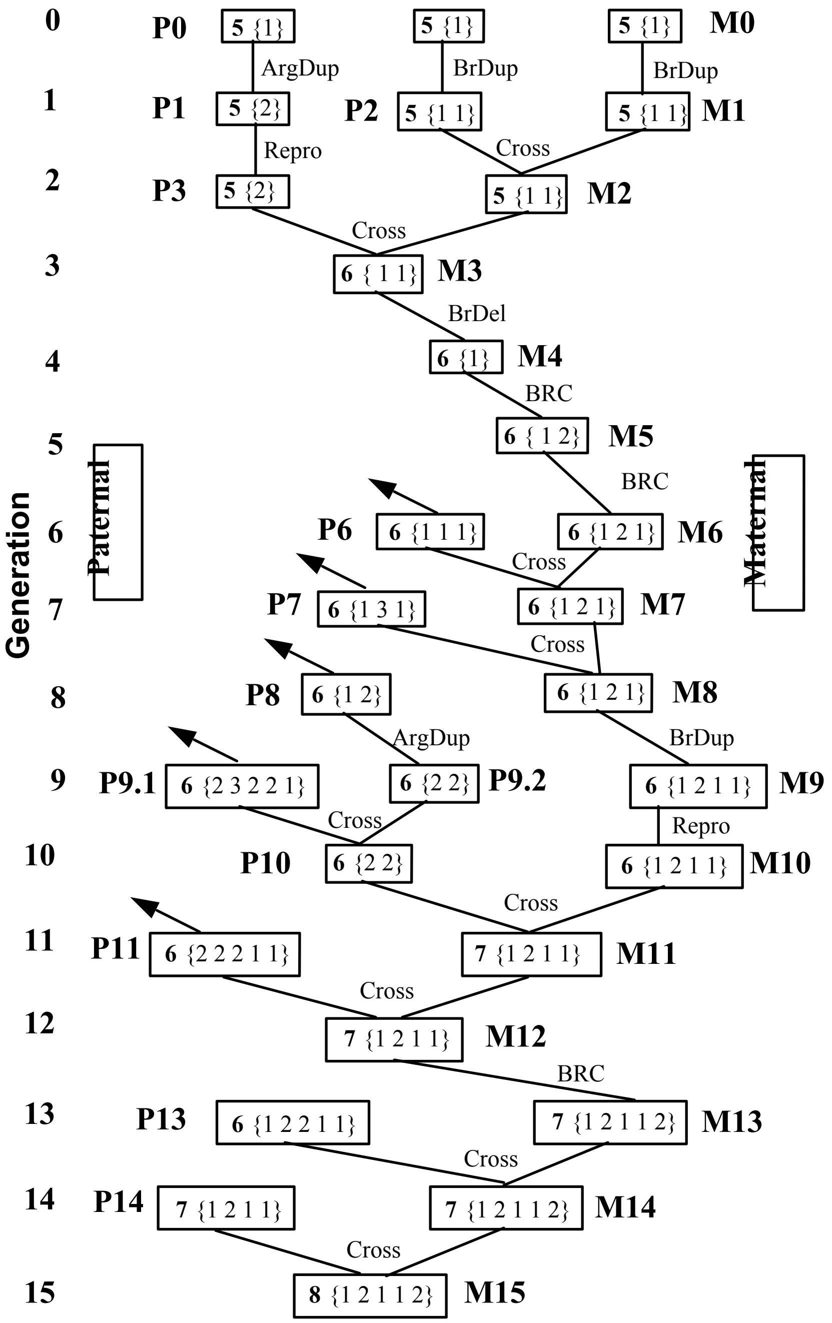
```
(progn (defun ADF0 (ARG0)
     (values (or (AND (NAND ARG0 ARG0) (or
      ARG0 ARG0)) (NOR (nor ARG0 ARG0) (AND
      ARG0 ARG0)))))

        (values (nor (AND D0(nor D2 D1))
         (AND (AND D2 D1)))))
```

EVEN-3-PARITY PROBLEM – BEST-OF-GENERATION 10 - ARGUMENT MAP OF {2, 2, 3, 2, 2} – 100%-CORRECT SOLUTION

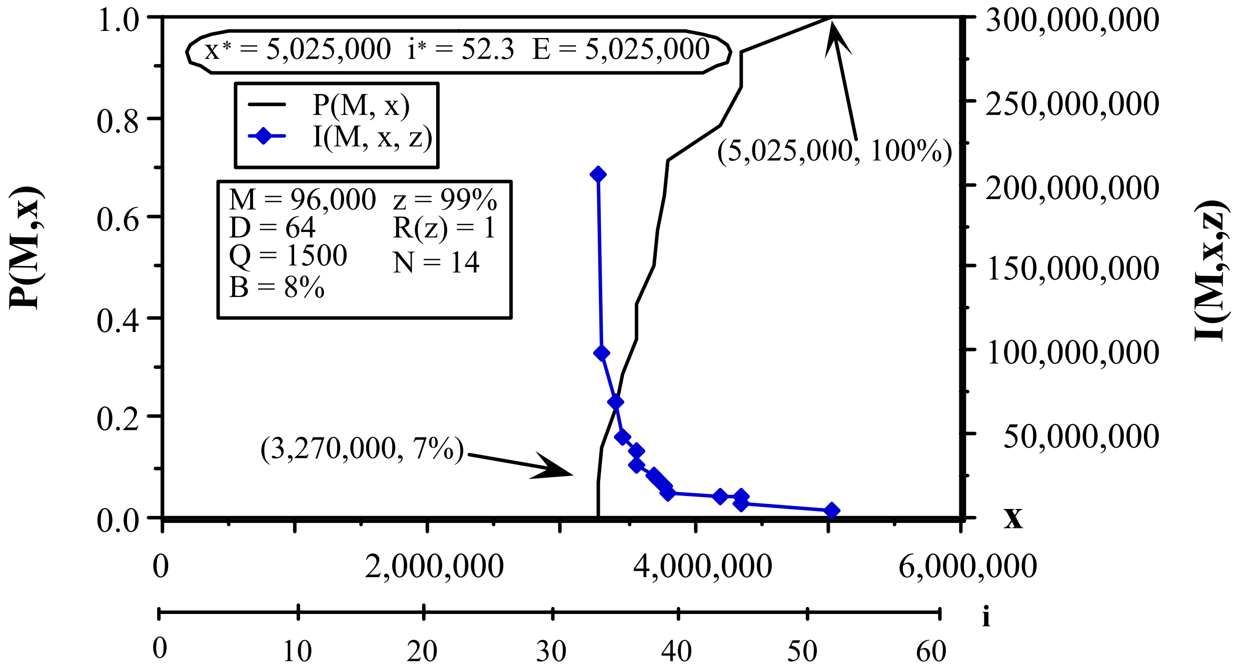
```
(progn (defun ADF0 (ARG0 ARG1)
     (values ((or (AND (NAND ARG0 ARG0) (or
      ARG1 ARG0)) (nor (nor ARG1 ARG0) (AND
      ARG0 ARG1))))))
     (defun ADF1 (ARG0 ARG1)
     (values ((or (AND ARG0 ARG1) (nor ARG0
      ARG1))))
     (defun ADF2 (ARG0 ARG1 ARG2)
     (values ((AND ARG1 (nor ARG0 ARG2))))
     (defun ADF3 (ARG0 ARG1)
     (values (ARG0))
     (defun ADF4 (ARG0 ARG1)
     (values ((or (AND ARG0 ARG1) (nor ARG0
      ARG1))))

       (values (nor (ADF4 D0(ADF1 D2 D1))
        (AND (ADF1 D2 D1) D0))))
```

RESULT-PRODUCING BRANCH OF THE 100%-CORRECT BEST-OF-RUN INDIVIDUAL FROM GENERATION 10 IS EQUIVALENT TO ...

```
(NOR  (even-2-parity D0(even-2-
parity D2 D1))
      (AND (even-2-parity D2
D1) D0))
```

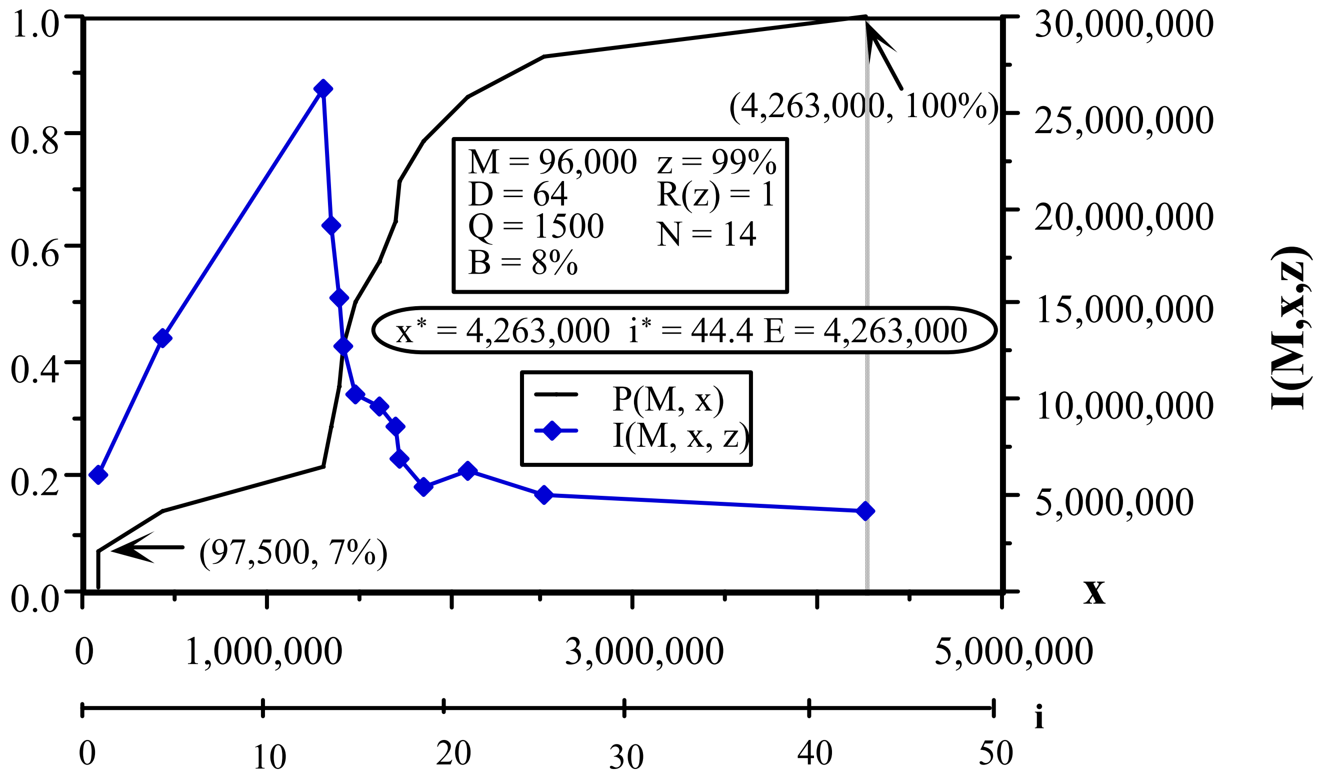

Generation
Paternal
Maternal
0
P0
5 {1}
5 {1}
5 {1}
M0
ArgDup
BrDup
BrDup
1
P1
5 {2}
P2
5 {1 1}
5 {1 1}
M1
Repro
Cross
2
P3
5 {2}
5 {1 1}
M2
Cross
3
6 { 1 1}
M3
BrDel
4
6 {1}
M4
BRC
5
6 { 1 2}
M5
BRC
6
P6
6 {1 1 1}
6 {1 2 1}
M6
Cross
7
P7
6 {1 3 1}
6 {1 2 1}
M7
Cross
8
P8
6 {1 2}
6 {1 2 1}
M8
ArgDup
BrDup
9
P9.1
6 {2 3 2 2 1}
6 {2 2}
P9.2
6 {1 2 1 1}
M9
Cross
Repro
10
P10
6 {2 2}
6 {1 2 1 1}
M10
Cross
11
P11
6 {2 2 2 1 1}
7 {1 2 1 1}
M11
Cross
12
7 {1 2 1 1}
M12
BRC
13
P13
6 {1 2 2 1 1}
7 {1 2 1 1 2}
M13
Cross
14
P14
7 {1 2 1 1}
7 {1 2 1 1 2}
M14
Cross
15
8 {1 2 1 1 2}
M15

COMPARISON OF FIVE APPROACHES TO SOLVING THE EVEN-5-PARITY PROBLEM

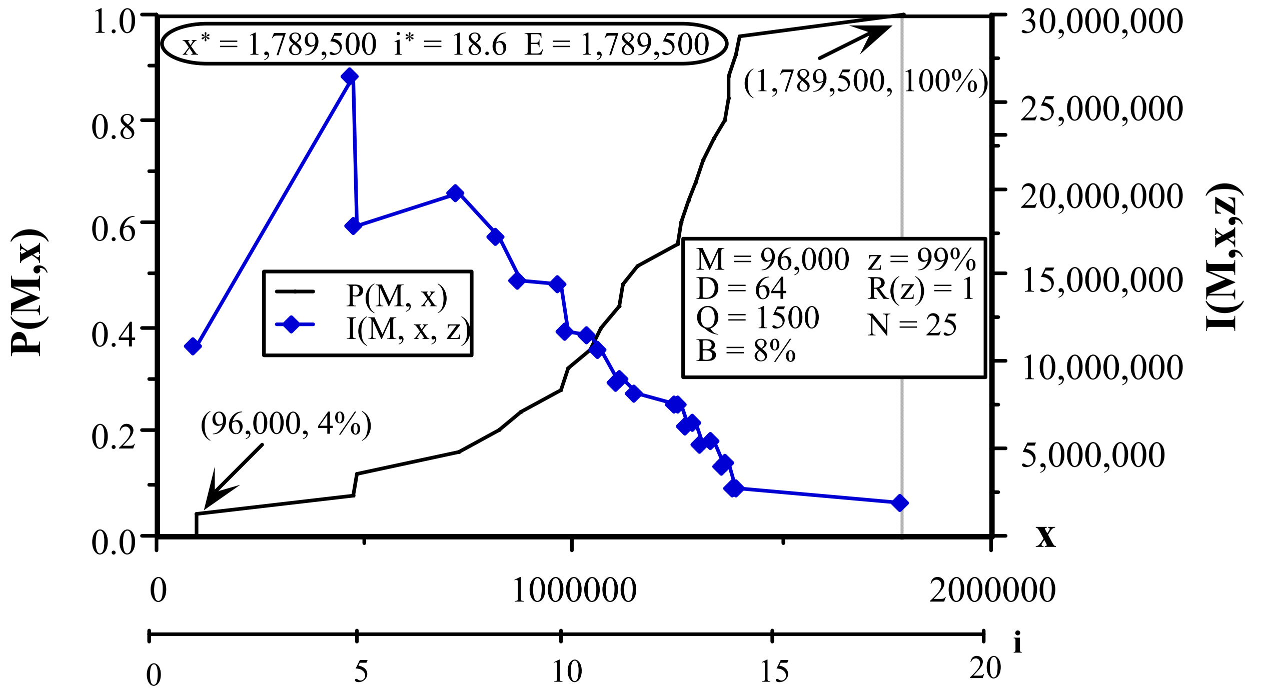


PERFORMANCE CURVES WITHOUT AUTOMATICALLY DEFINED FUNCTIONS (APPROACH A)

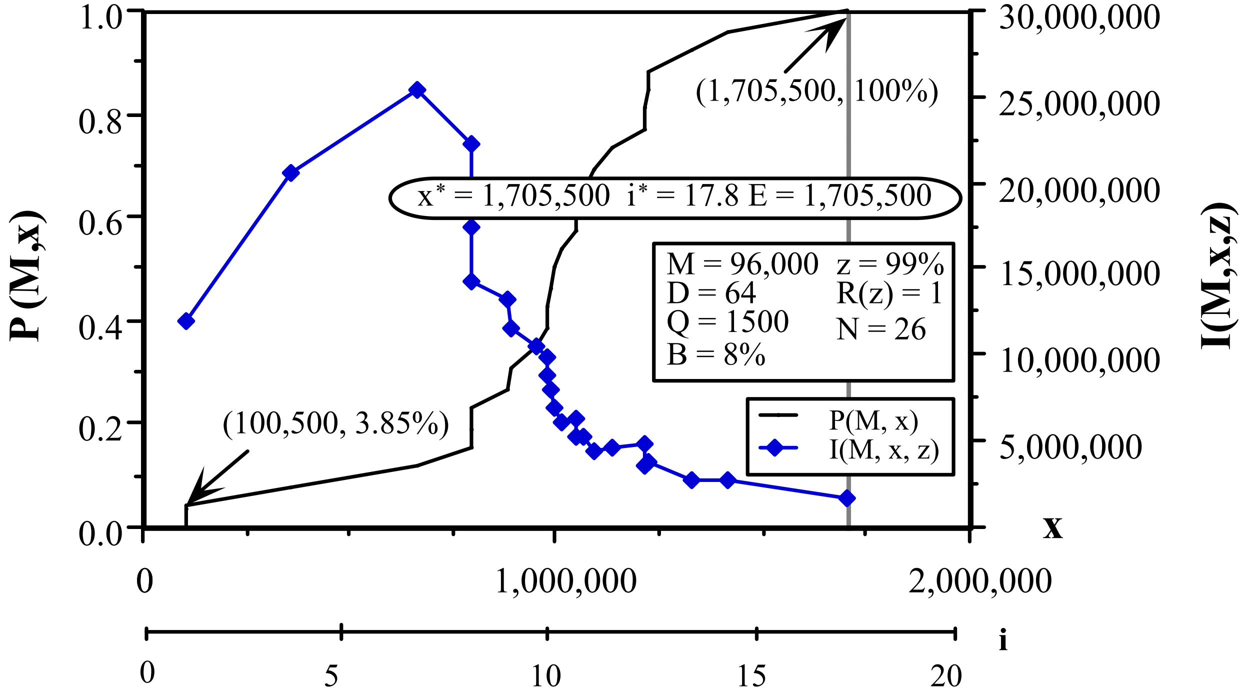
PERFORMANCE CURVES WITH EVOLUTIONARY SELECTION OF THE ARCHITECTURE (APPROACH B)



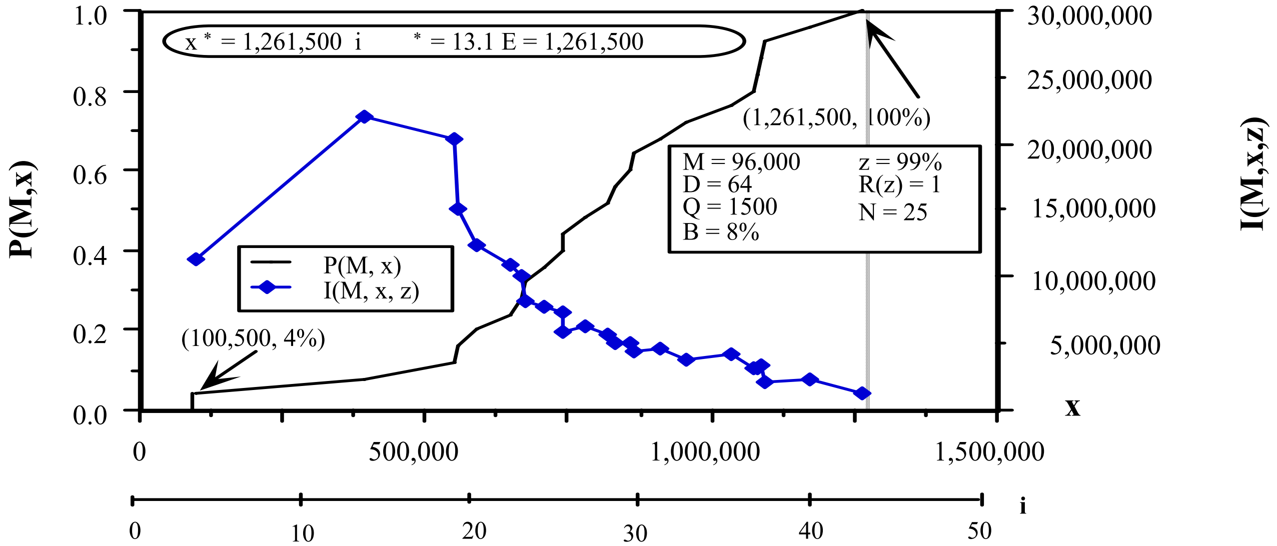
PERFORMANCE CURVES FOR THE PROBLEM OF SYMBOLIC REGRESSION OF THE EVEN-5-PARITY FUNCTION USING THE ARCHITECTURE-ALTERING OPERATIONS (APPROACH C)



PERFORMANCE CURVES FOR THE FIXED {3, 2} ARCHITECTURE AND POINT TYPING (APPROACH D)



PERFORMANCE CURVES FOR THE FIXED {3, 2} ARCHITECTURE AND BRANCH TYPING (APPROACH E).



Comparison of five approaches to solving the even-5-parity problem

Approach	**Number of runs**	**Computational effort *E***	**Wallclock time $W(M,t,z)$**	**Average structural complexity $\bar{S}$**
A - No ADFs	**14**	**5,025,000**	**36,950**	**469.1**
B - Evolutionary selection of the architecture	**14**	**4,263,000**	**66,667**	**180.9**
C - Architecture-altering operations	**25**	**1,789,500**	**13,594**	**88.8**
D - Fixed architecture with point typing	**25**	**1,705,500**	**14,088**	**130.0**
E- Fixed architecture with branch typing	**25**	**1,261,500**	**6,481**	**112.2**

- $E(\mathrm{A}) > E(\mathrm{B}) > E(\mathrm{C}) > E(\mathrm{D}) > E(\mathrm{E})$
- $W(\mathrm{A}) > W(\mathrm{C}) > W(\mathrm{E})$
- $\bar{S}(\mathrm{A}) > \bar{S}(\mathrm{B, C, D, E})$
- $\bar{S}(\mathrm{A, B, D, E}) > \bar{S}(\mathrm{C})$