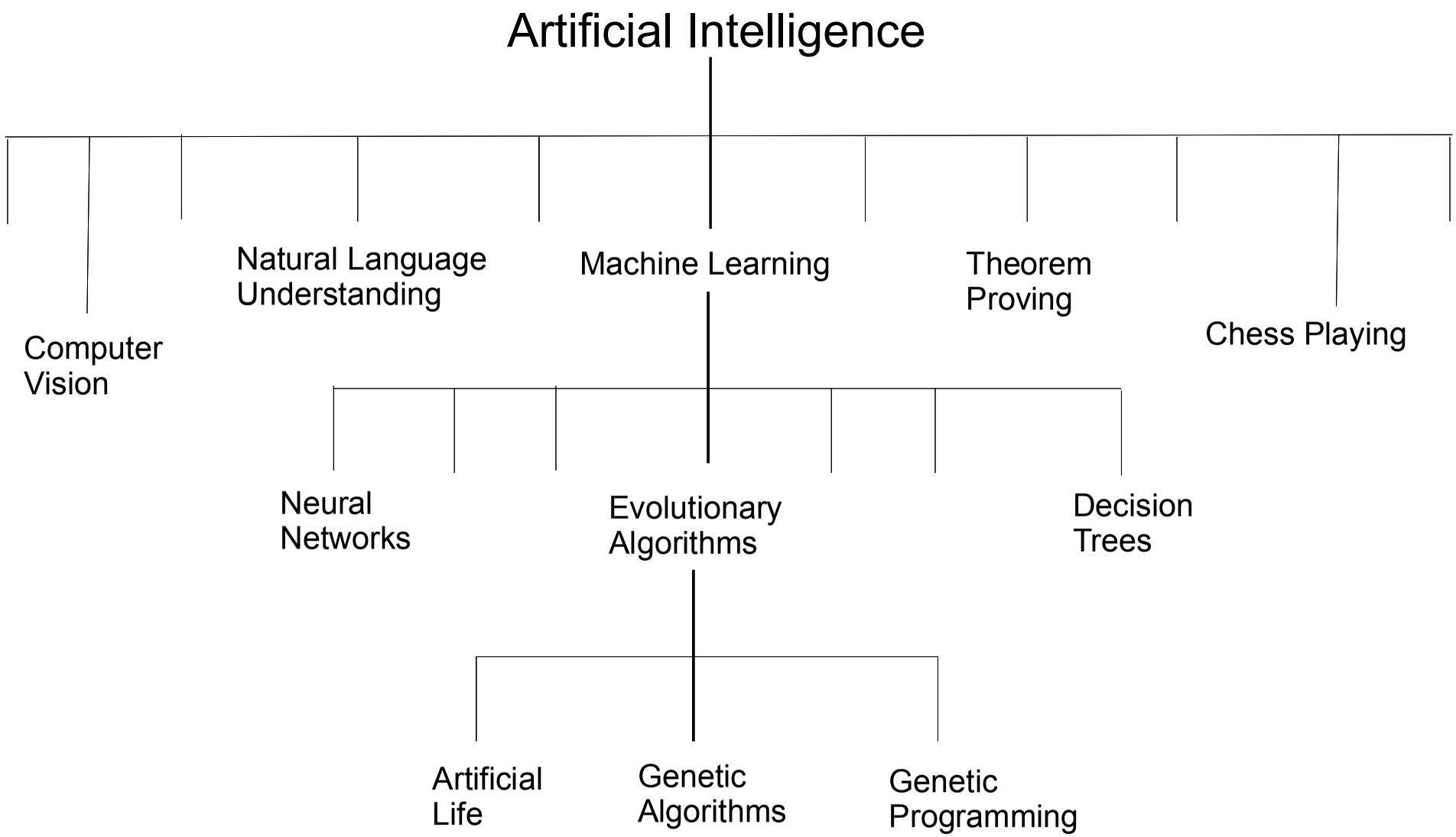
Using Evolutionary Algorithms To Solve Hard Problems

(ralph.morelli@trincoll.edu)

Outline

- What are Genetic Algorithms.
- Some examples from different domains.
- Themes
  - GA: inspired by natural selection.
  - Simple model: Random variation plus selection.
  - Complexity: emergent property.
- Implications for secularism? (discussion)
  - Watches without watch makers.

What are GAs?


Artificial Intelligence
Computer Vision
Natural Language Understanding
Machine Learning
Theorem Proving
Chess Playing
Neural Networks
Evolutionary Algorithms
Decision Trees
Artificial Life
Genetic Algorithms
Genetic Programming

GAs Competitive with Humans

- AI researchers keep score (Turing Test)
- 1997: Deep blue beat Kasparov.
- http://www.genetic-programming.com/humancompetitive.html
- Some examples from Genetic Programming (out of 36)
  - Creation of quantum algorithm for 'early promise' problem.
  - Creation of quantum algorithm for Grover's DB search.
  - Rediscovery of Cauer elliptical topology for filters.
  - Synthesis of a NAND circuit.
  - Rediscovery of negative feedback.

Solving Short Cryptograms

tmlk sceq adqeq e dj blekb sl qclv ylt clv d atkrc ln apebcs yltkb nlhgq wew nekw d rcdjmelk; d jdk vesc alyq dkw bephq ln ceq lvk.

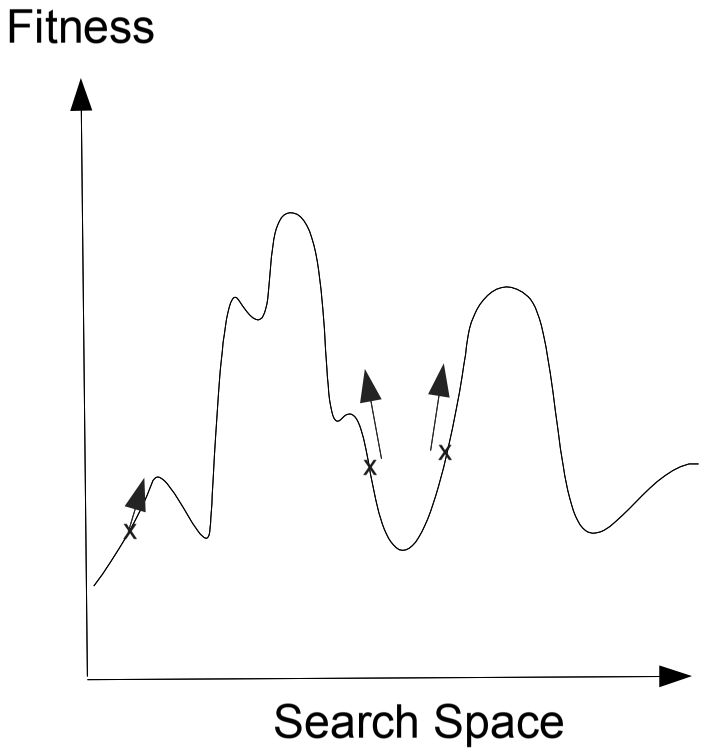
- Short cryptograms are 'hard'.
  - Simple substitution cipher: a/e, b/g, c/t, ..., x/c, y/m, z/j
  - Key: A permutation of the alphabet, a..z.

    ZWUSQOMKIGECABDFHJLNPRTVXY
  - Possible keys: 26! = 4.0329 * 10^{26}
  - Unsolvable if too short
    - Theoretical minimum length ~ 28 characters
    - To we or not to we that is the question
    - Xabcx = River = Raver = Rover = Saves = David ...

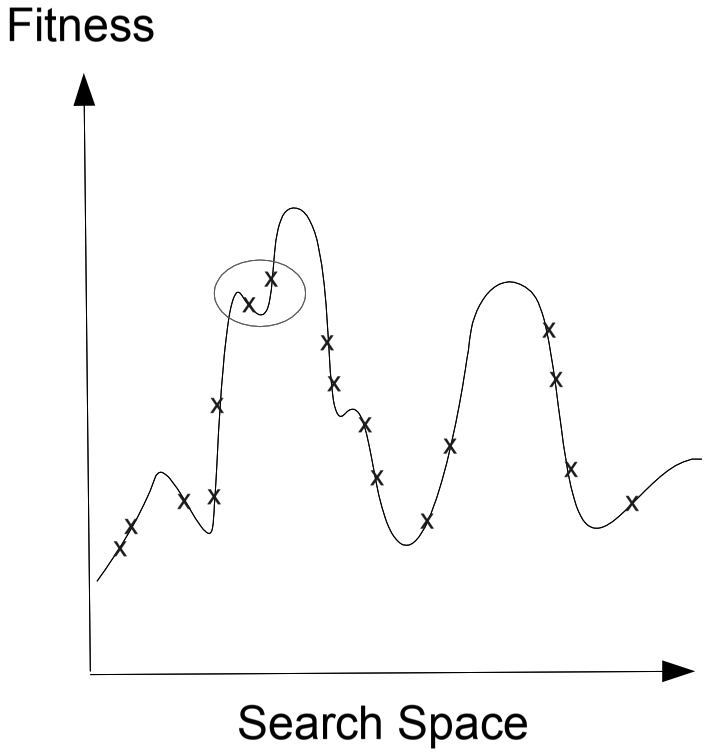
Searching a Problem Space

- *Algorithmic search*
  - Try every key (brute force).
  - Guaranteed to work (theoretically).
  - Intractable: $10^{26}/10^{9} = 10^{17}$ seconds $\approx 10^{9}$ years
- *Heuristic search*
  - Use a *rule of thumb* to reduce search space.
  - Not guaranteed to succeed.
  - E.G. Traveling salesman *nearest neighbor* heuristic.
  - E.G. *Hill climbing* optimization.
  - E.G. *Genetic algorithm*.

Hill Climbing vs GA

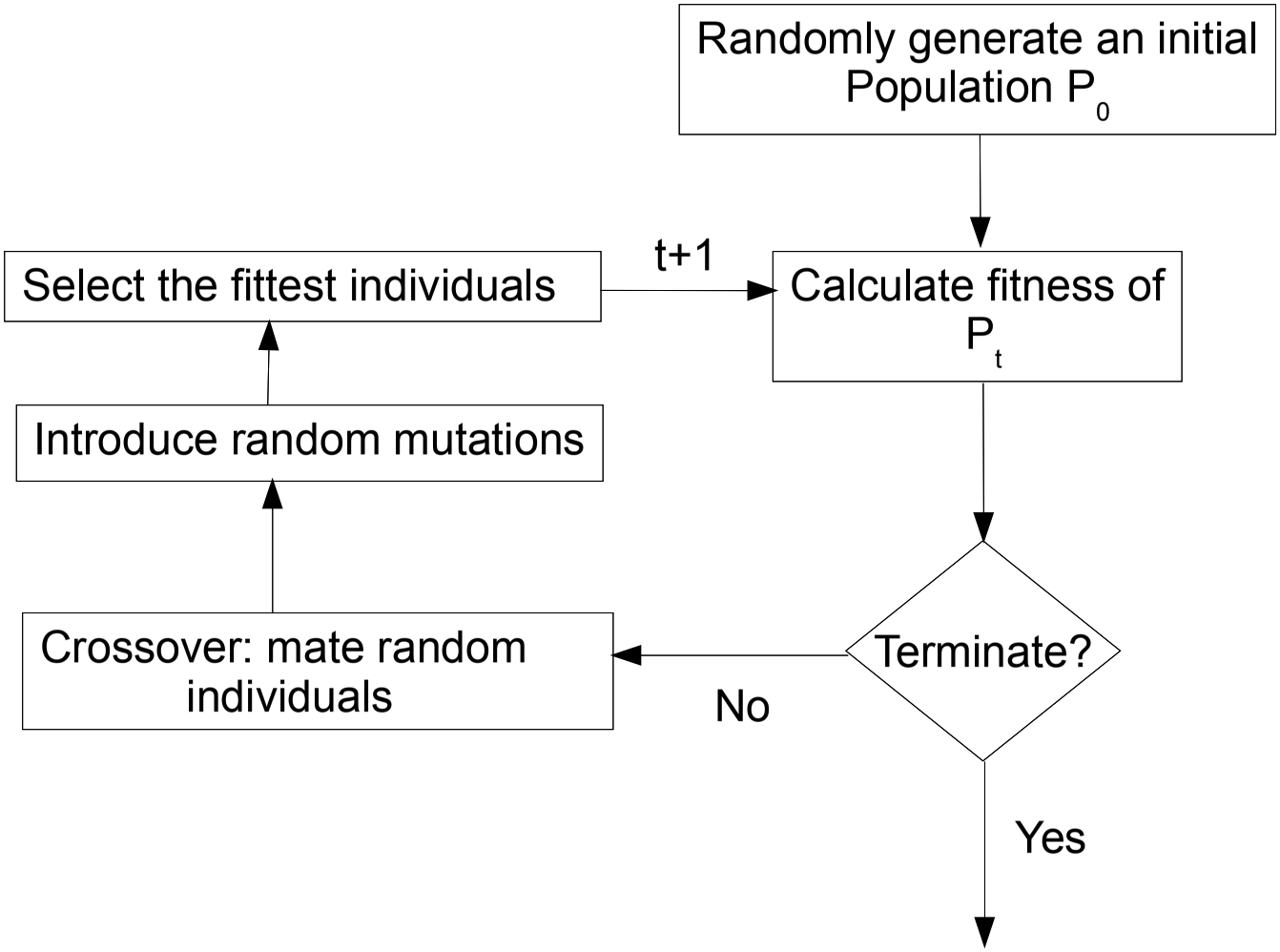


Hill Climber

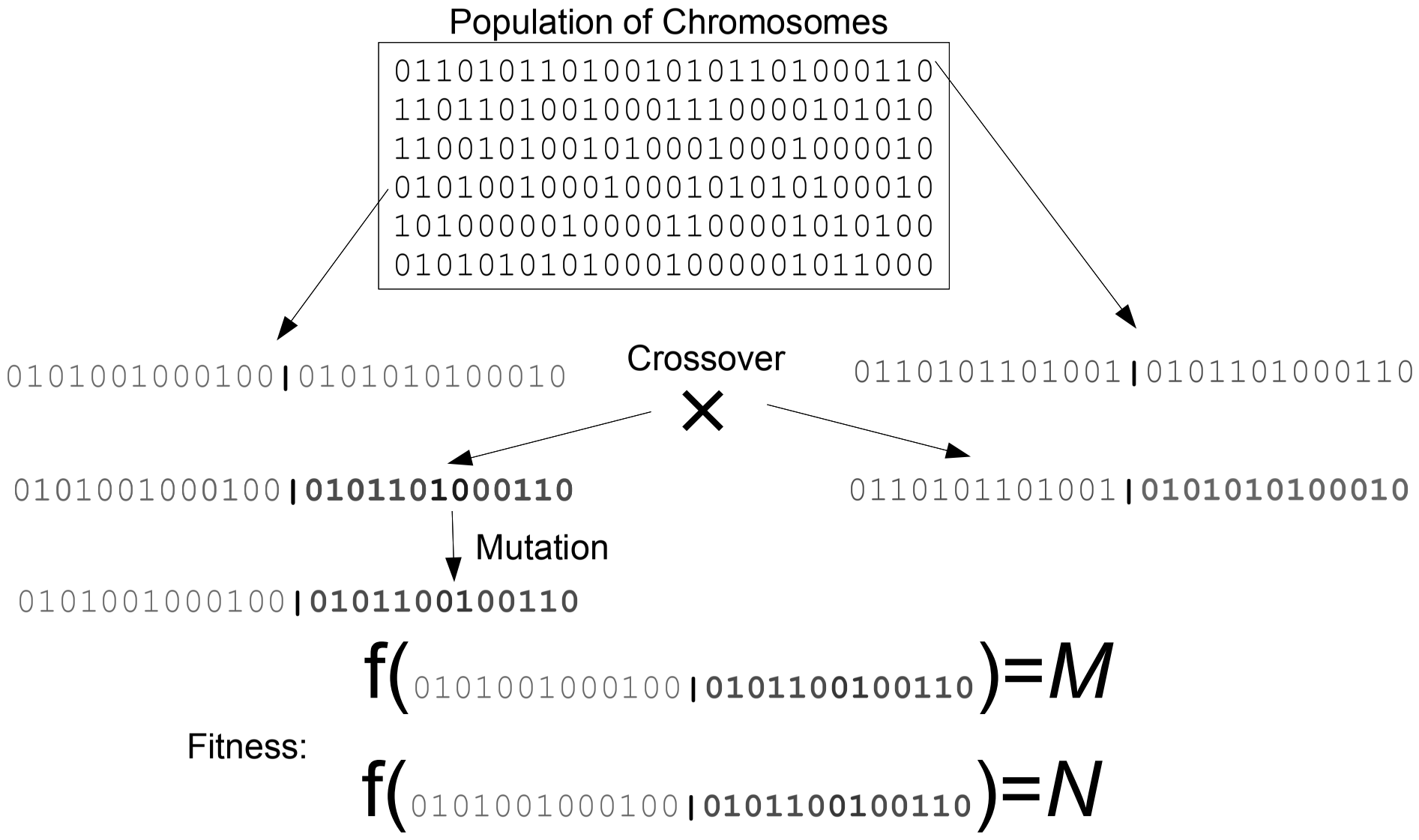


Genetic Algorithm

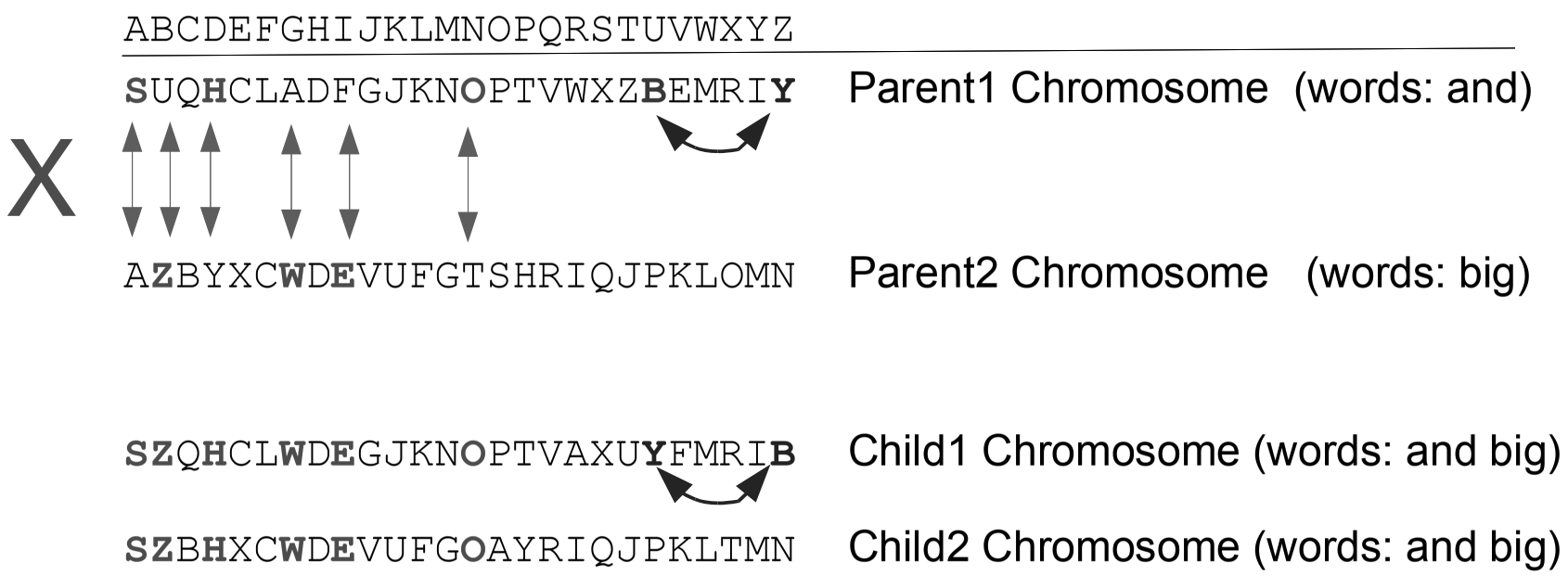
The Genetic Algorithm



GA Basics


Population of Chromosomes
0110101101001010110100011 0
1101101001000111000010101 0
1100101001010001000100001 0
0101001000100010101010001 0
1010000010000110000101010 0
0101010101000100000101100 0
0101001000100|0101010100010
Crossover
0110101101001|0101101000110
0101001000100|0101101000110
0110101101001|0101010100010
Mutation
0101001000100|0101100100110
f(0101001000100|0101100100110)=M
Fitness:
f(0101001000100|0101100100110)=N

Cryptogram: Crossover and Mutation



f(SZBHXCWDEVUFGOAYRIQJPKLTMN)= "... gx … and nycd ... big ..."

Results

Parameter	Setting
Population size	512
Seed dictionary size	50
Fitness dictionary size	3500
Crossover rate	0.9
Mutation rate	0.2

msg	nToks	nWds	%Found	nGens	nTrials
1	24	19	97.4	21.1	50
2	33	22	99.5	16.2	50
3	23	14	61.5	46.9	50
4	23	12	58.6	52.4	50
5	26	17	95.7	16.1	50
6	24	22	93.1	57.8	50
7	25	16	74.0	46.2	50

Table 3: Performance results for seven cryptograms.

Multiple Sequence Alignment

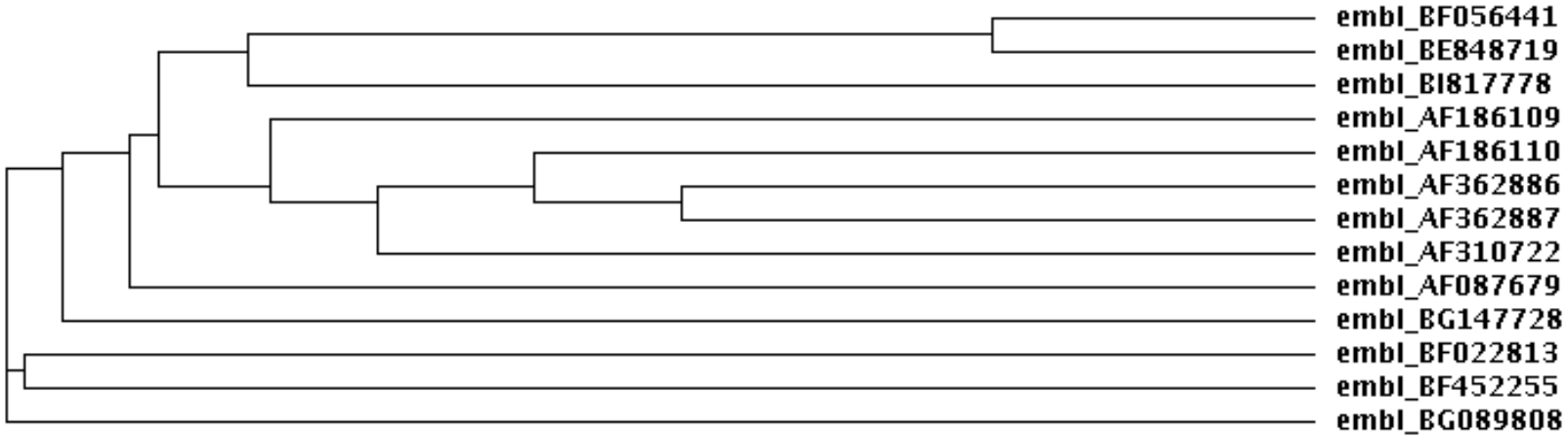
- Fundamental molecular biology problem:
  - Identify common structure in a string of nucleotides (DNA, RNA) or amino acids (in proteins).
- For “deciphering”
  - Evolutionary history
  - Phylogenetic relationships among organisms.
- Very hard ~ on the order of m^n operations for n sequences of length m.

Clustal-W Example

- From a set of DNA sequences such as these:

```
embl_BF022813      ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 86
embl_BF452255      ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 83
embl_BG089808      ATGGCCGGCCTCAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 94
embl_BG147728      ----------CAACTCACTGGAGGCAGTGAAGCGCAAGATCCAGGCCCTGCAGCAGCAG 49
embl_AF087679      ATGGCCGGCCTCAACTCCCTGGAGGCGGTGAAACGCAAGATCCAGGCCCTGCAGCAGCAG 60
embl_AF362886      ------------------------------------------------------------
embl_AF362887      --------------------------------CGAGAAGTTGAGGGAGAAAGGCGGGCC 27
embl_AF186110      ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 285
embl_AF310722      ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 360
embl_AF186109      ATGAAGGATGAGGAGAAGATGGAGATTCAGGAGATGCAGCTCAAAGAGGCCAAGCACATT 357
embl_BI817778      ATGTCGGGTGGCAGTTCCATCGATGCGGTGAAGAAGAAGATCCAGAGCCTTCAGCAGGTG 137
embl_BF056441      GTATTCATAATTTGCTTGACATTTCCAGCAAAGCGAAGATGGCAATAACAAAAGGAACTT 271
embl_BE848719      -TGTTACCAATCTGCTTGGCATTTCCTGCAAGGTGGAAACC-TGGTAATAAGCGGAACTT 58
```

- Compute a phylogram such as this:



The Algorithm

- The chromosomes were candidate alignments, represented in a matrix with *n* rows.
- Create a random population of candidate alignments.
- For each candidate, apply *variation* operators to derive a child candidate.
- Apply a selection operator (fitness test) to generate the next generation.
- Repeat until 200 generations, no change in 100 generations, or number of gaps fell below a certain threshold.

Evolutionary Programming

- Chellapilla and Fogel's EP is comparable to Clustal-W.

Table 1: Data sets used for testing the proposed evolutionary programming procedure for multiple sequence alignment. Information regarding the data sets is provided in the Appendix.

Data Set	Number of Sequences	Mean Sequence Length in nucleotides (min,max)	Percentage of matched columns based on the best alignment using ClustalW	Number of Matched Columns in ClustalW solution	Number of Matched Columns in EP solution	Number of Generations	EP Score
1. S1 (Zhang, 1997)	10	211.9 (211, 212)	93.39	198	198	200	4082
2. 16S rRNA	8	457.0 (457, 457)	98.25	449	449	400	7233
3. Histone H3	21	122.0 (122, 122)	89.34	109	109	160	4766
4. Histone H3-H4 Intergenic Region	21	333.4 (322, 346)	27.41	91	102	180	7033

- EP Advantages
  - Flexibility in fitness functions.
  - Tackles harder and longer sequences.
  - Outperforms Clustal-W for low-similarity sequences.

Thank You!