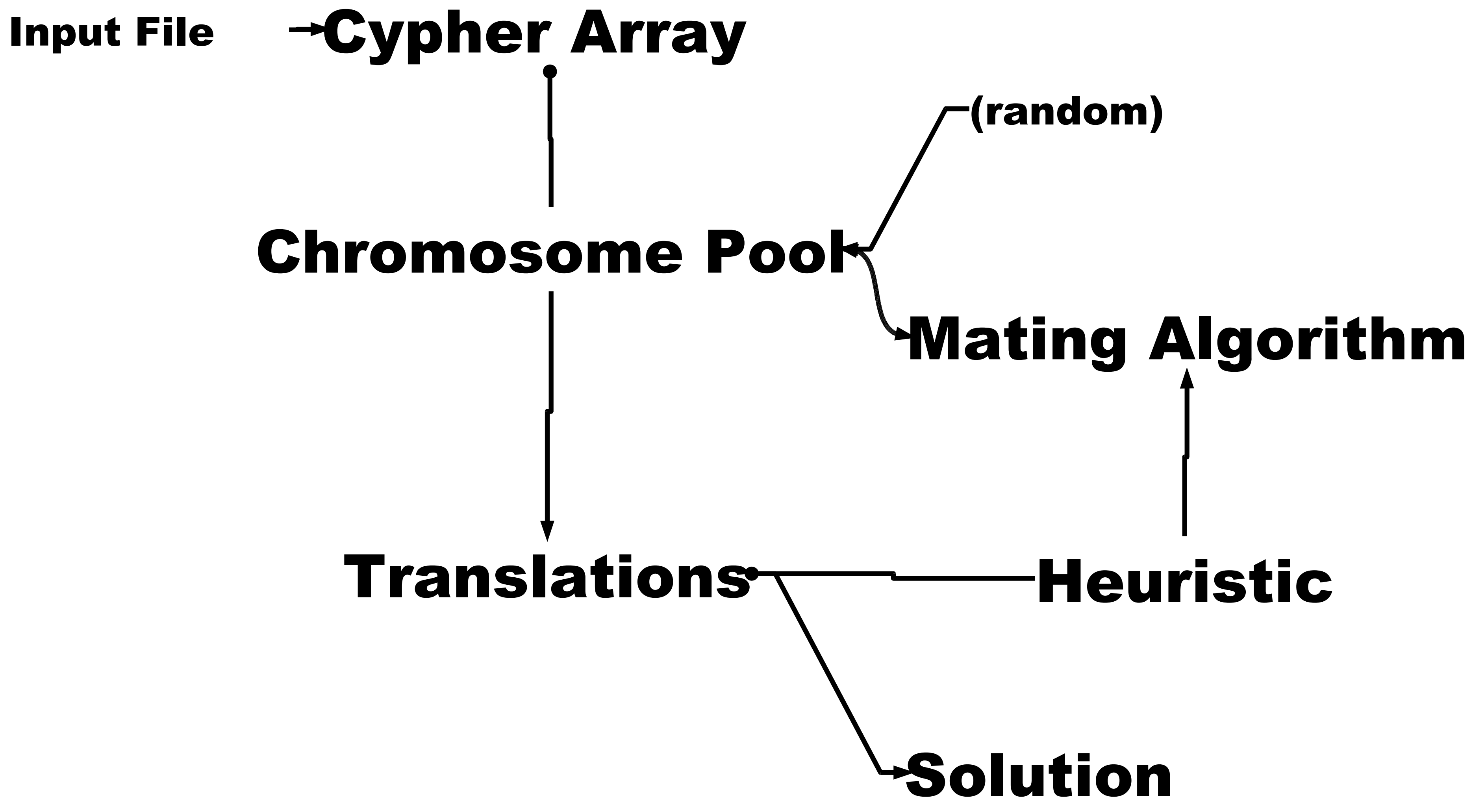
OPTIMIZING GENETIC ALGORITHMS FOR USE IN DECODING CYPHERS

Keith Diggs — 2004-2005

Computer Systems Research Lab

ABSTRACT

Over the past several years, genetic algorithms have come into wide use because of their ability to find good solutions to computing problems very quickly. They imitate nature by crossing over strings of information represented as chromosomes, with preference given to the more fit solutions produced. They hold great promise in the field of cryptology, where they may be used to quickly find good partial solutions, thus eliminating much of the intense manual labor that goes into identifying initial coding schemes.



Input File

Cypher Array

(random)

Chromosome Pool

Mating Algorithm

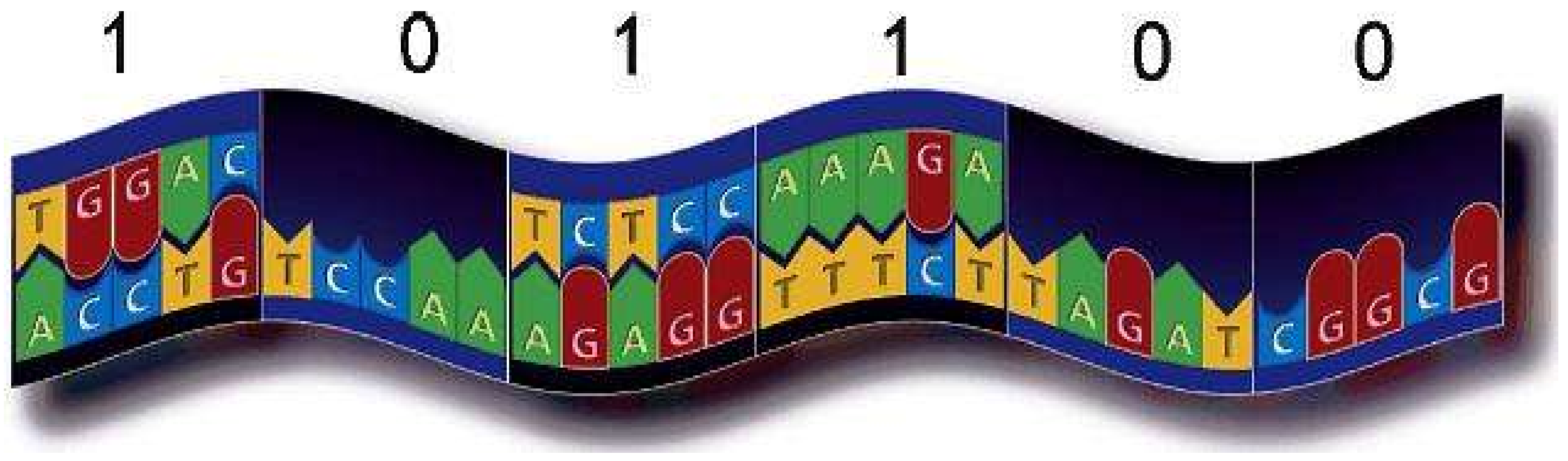
Translations

Heuristic

Solution

PROCESS

The basic data structure of the program consists of two arrays: one to contain the numbers of the actual cypher, another to hold the chromosomes (possible keys to translation). The flow chart at left represents the program's process: the cypher array is ran past the chromosome pool to create provisional translations of the cypher, which are evaluated by the heuristic to determine the fitness of the chromosome that helped create the translation. This dictates the way in which the chromosomes are mated (the stronger chromosomes will have a greater chance of mating). The hope is that this will produce an intelligible and correct solution to the cypher.



BACKGROUND

The Beale Cypher is one of the most famous unsolved puzzles in cryptography. Basically, 100 years ago, Beale buried treasure in a lake in Bedford County, near Roanoke. Three letters were written, encoded and given to a friend. Beale subsequently travelled west and never returned.

Later, one of the 3 letters was deciphered. Beale had used a simple encoding algorithm. His letters consisted of a large list of numbers. These numbers corresponded to the first 480 words in the American Declaration of Independence. For example, if the document starts "The quick brown fox..." then 3 would represent the letter 'b' and so forth.

So, one of the letters was succesfully deciphered, but to this day the other two remain unbreakable.

Source: Matthews, James, *The Beale Cypher,* <http://www.generation5.org/content/2003/beale.asp>

THE HEURISTIC

The heuristic algorithm is based on two components: a character-frequency analysis and a word search. Chromosomes whose relative character frequencies closely match that of the English language will be assigned higher fitness values, as will chromosomes that are found to contain a limited set of actual English words.