

15 May - 9 June 1995

Genetic Algorithms for Predicting Biomolecular Structure

V. Sundararajan

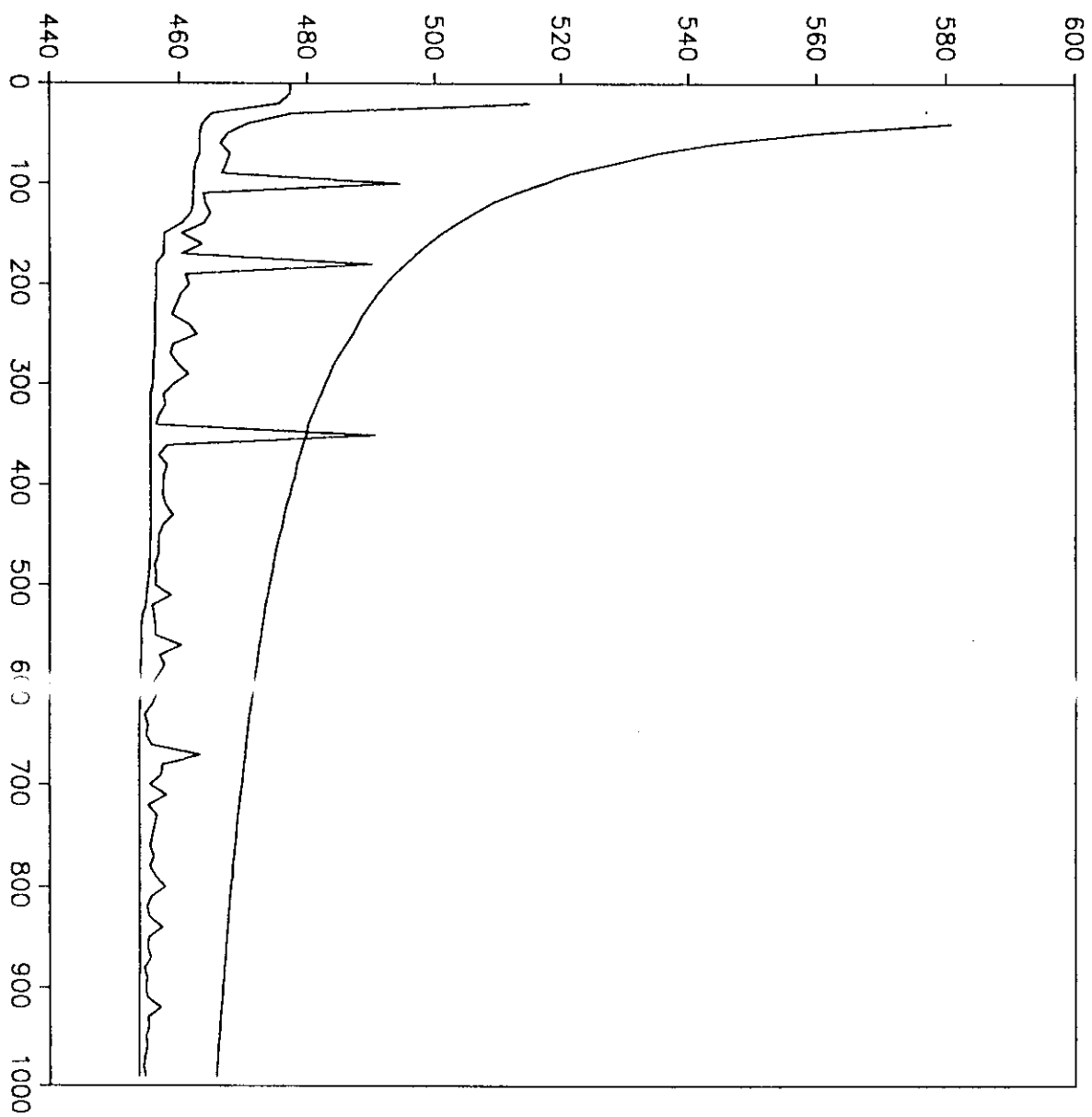
Centre for Development of Advanced Computing
Pune, India

GENETIC ALGORITHMS FOR PREDICTING BIOMOLECULAR STRUCTURE

V. SUNDARARAJAN

CENTRE FOR DEVELOPMENT
OF ADVANCED COMPUTING
PUNE, INDIA.

Prof. A-S. KOLASKAR
DEPT. BIOTECHNOLOGY
UNIV. OF PUNE
PUNE 411007 INDIA.



PREDICTING STRUCTURE OF BIOMOLECULES

Energy Optimisation Problem

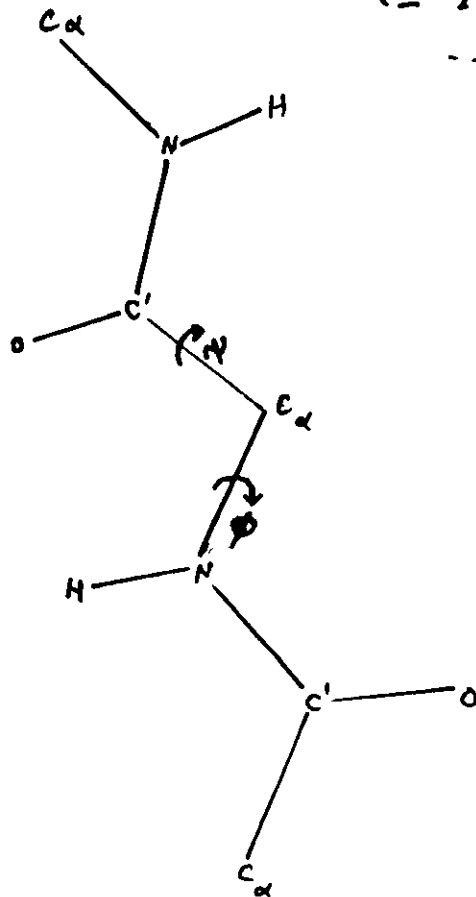
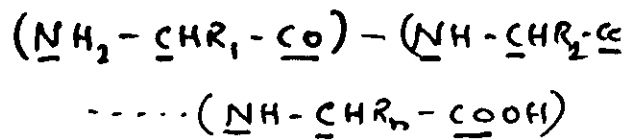
Atomic Positions

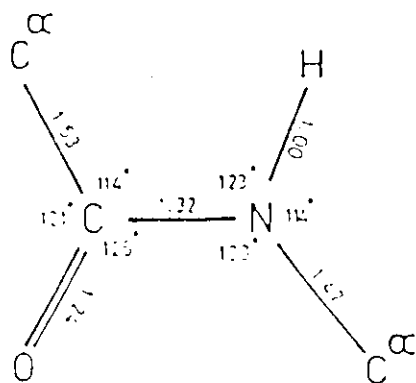
or

bond lengths, bond angles, Dihedral
angles

Typically 1000 to 10,000 atoms.

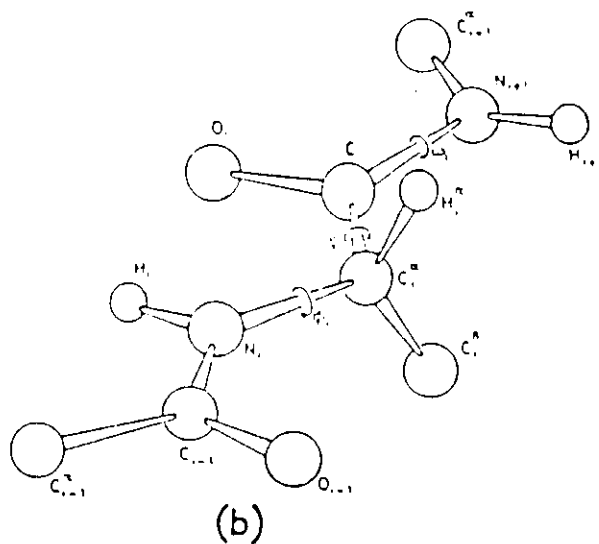
PROTEINS → LINKED AMINO ACID RESIDUES



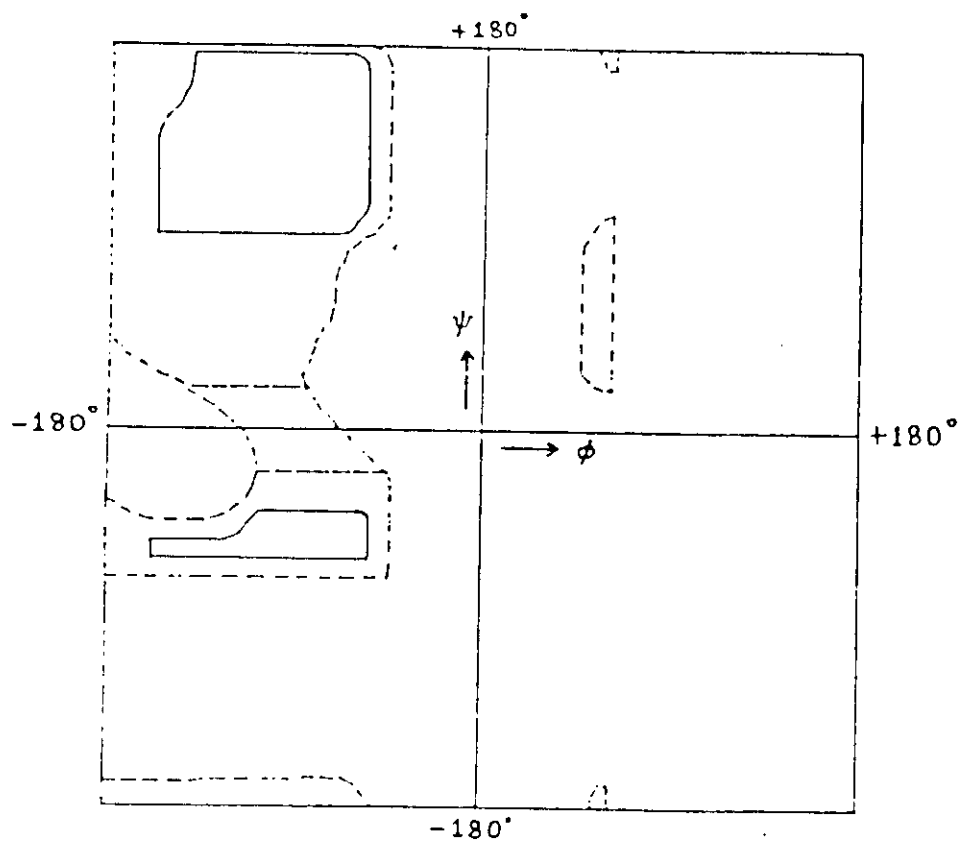


trans

(a)

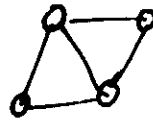
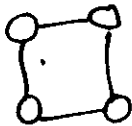
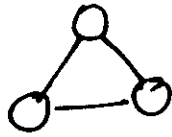
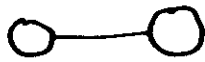


(b)



(c)

WHY OPTIMISATION ?



METHODS :

1. CALCULUS :
 - STEEPEST DESCENT
 - CONJUGATE GRADIENT
 - VARIABLE METRIC

2. ENUMERATIVE :

3. RANDOMISED

GENETIC ALGORITHMS

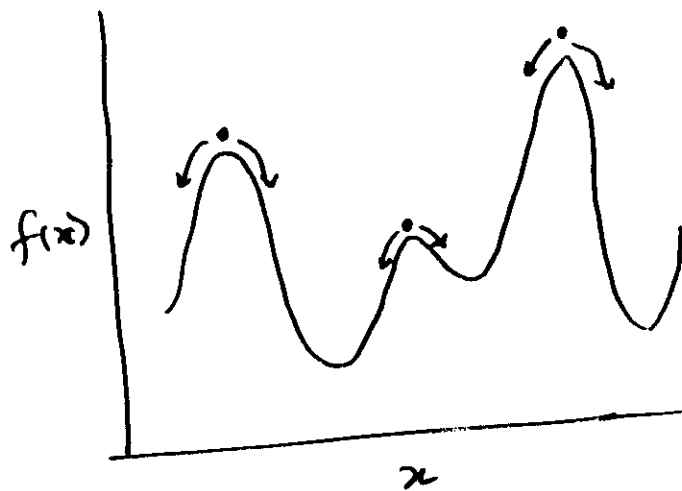
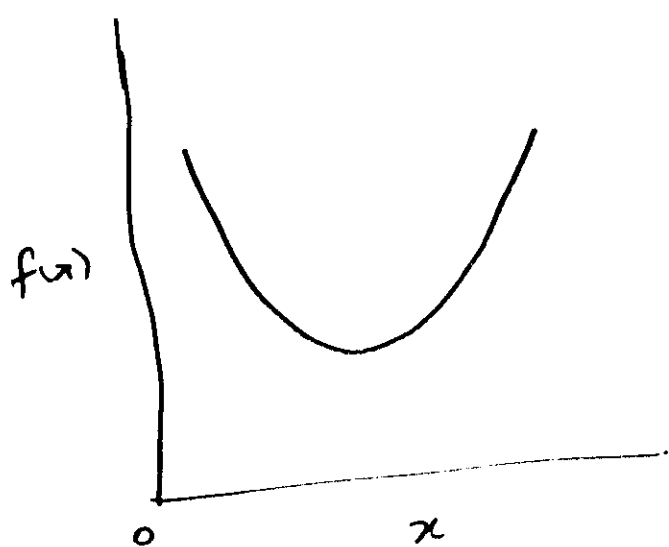
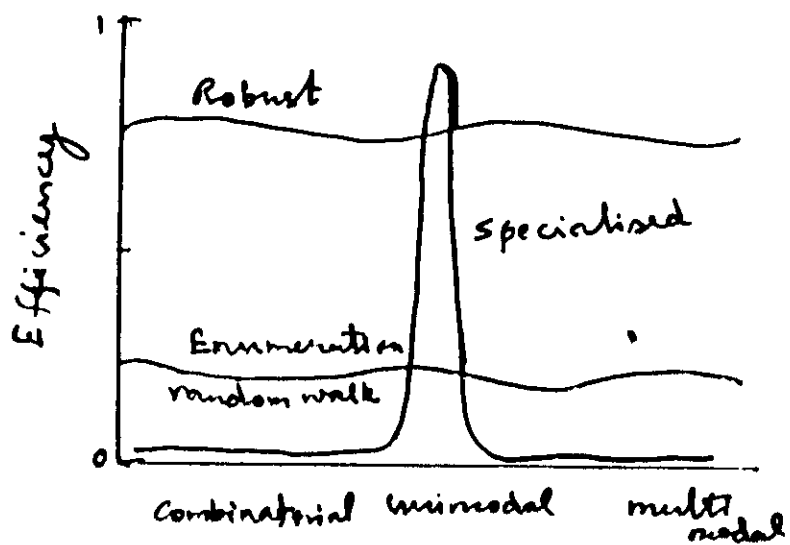
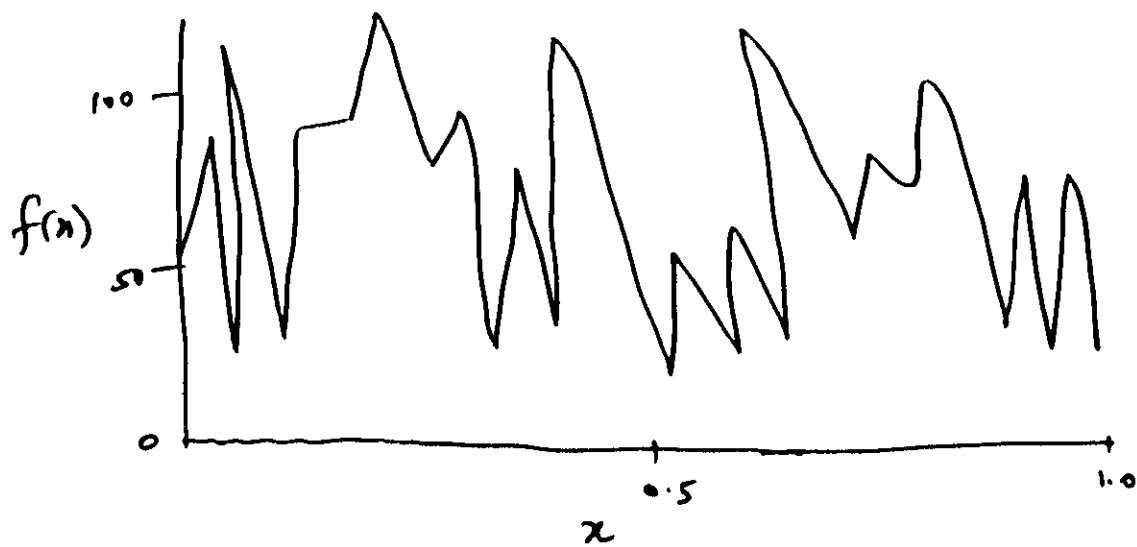
SIMULATED ANNEALING

MOLECULAR DYNAMICS .

GOAL

PROCESS → IMPROVEMENT

OPTIMAL POINT



GENETIC ALGORITHM : FEATURES

(Holland, 1975)

1. DARWINIAN THEORY OF EVOLUTION OF BIOLOGICAL SYSTEMS

GA EVOLVES SOLUTIONS.

2. GA IS BASED ON POPULATION GENETICS.

POPULATION OF GENOTYPES STORED IN MEMORY,
REPRODUCTION AND VARIATIONS BY PROCESSES SIMILAR
TO MUTATION AND CROSS-OVER.

ACTUALLY

3. OPERATES ON BINARY STRINGS.

INTENDED

4. a) TO ABSTRACT AND RIGOROUSLY EXPLAIN THE
ADAPTIVE PROCESSES OF NATURAL SYSTEMS

b) TO DESIGN ARTIFICIAL SYSTEMS SOFTWARE

ULTIMATE GOAL

5. ROBUSTNESS, EFFICIENCY AND FLEXIBILITY OF
BIOLOGICAL SYSTEMS HAVE TO BE ADAPTED.

FEATURES OF SELF-REPAIR, SELF-GUIDANCE
AND REPRODUCTION.

— DREAM !!!

- SEARCH FROM A POPULATION OF POINTS
- NO DERIVATIVES
- LOCATE GLOBAL EXTREMUM
- INHERENT PARALLELISM
- TOO MANY FUNCTION EVALUATIONS

MATHEMATICAL BASIS

SCHEMA - SIMILARITY TEMPLATE $\{1, 0, *\}$

SCHEMA THEOREM

$$m(s, t+1) \geq \left[m(s, t) \frac{f(s)}{f(\mu)} \right] \left\{ 1 - p_c \frac{\delta(s)}{l-1} - o(s)p_m \right\}$$

m - No. of representatives of schema s at time t .

$f(s)$ - average fitness of schema s .

$f(\mu)$ - overall average fitness of population

$\delta(s)$ - definite length

$o(s)$ - order of schema

l - length of string representing schema.

p_c - probability of crossover

p_m - " " mutation

High objective, short length $\delta(s)$ and
low order $o(s)$ schemata grow exponentially

Template $\rightarrow$ string of bits or "genes" $\{1, 0, *\}$

Schema $\rightarrow$ 1 0 * * 1 0 * 1

1 0 1 1 1 0 0 1

1 0 0 1 1 0 1 1

1 0 0 0 1 0 0 1

CHROMOSOMES

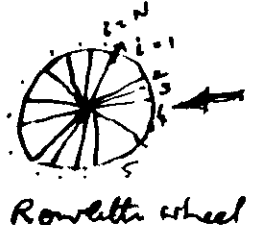
DEFINITE BITS

SCHEMATA $\rightarrow$ Plural of Schema

$O(s)$ - Order of schema - No. of Definite bits

$\delta(s)$ - defining length - Distance between right most * left most definite bits

$M(s, t)$ No. of examples of a particular Schema s contained in a population $A(t)$ at time t .

REPRODUCTION : Probability $\cdot \frac{f_i}{\sum_j f_j}$ 
COPIED ACCORDING TO FITNESS. Roulette wheel

CROSS OVER :

$$\text{Probability of Survival} = 1 - p_c \frac{\delta(s)}{l-1}$$

MUTATION :

$$\begin{aligned} \text{Probability of Survival} &= (1 - p_m)^{O(s)} \\ &\approx 1 - O(s)p_m \end{aligned}$$

ALGORITHM :

1. INITIAL POPULATION GENERATED RANDOMLY
2. SELECTION USING FITNESS
3. INHERITANCE - COPIES OF HIGH FITNESS INDIVIDUALS
4. MUTATION - FLIPPING INDIVIDUAL BITS
5. CROSS-OVER - EXCHANGE SUBSTRINGS TO GET OFFSPRING.

GENETIC OPERATORS (Crossover AND MUTATION)

APPLIED PROBABILISTICALLY TO PRODUCE NEW GENERATION OF INDIVIDUALS.

BY THIS CYCLE OF EVOLUTION , SELECTION AND GENETIC OPERATIONS ITERATED , OVERALL FITNESS OF POPULATION IMPROVES AND THE INDIVIDUALS REPRESENT "IMPROVED" SOLUTIONS TO WHATEVER PROBLEM POSED THROUGH FITNESS FUNCTION.

AN EXAMPLE

FINDING THE SQUARE ROOT OF A NUMBER '64'

USE 8 BIT SCHEMA TO REPRESENT 'x'

$$\text{FITNESS} = \text{INPUT} - x^2$$

START WITH RANDOM VALUES OF x

BEFORE CROSSOVER	Fitness
$\wedge$ 0 0 0 0 1 1 1 1	-161
1 1 1 1 0 0 1 0 $\wedge$	-58400

AFTER CROSSOVER	
1 1 1 1 0 1 1 1	-60945
0 0 0 0 1 0 1 0	-36
↑	
mutation at this bit	
↓	
0 0 0 0 1 0 0 0	0

8 is the answer

GAS for POLYPEPTIDE CONFORMATIONS :

1. POPULATION OF 20-50 CONFORMATIONS
2. ENERGY VALUE AS THE FITNESS
(LOOK FOR LOWER ENERGY) $E = E_{nb} + E_{es} + E_{tor} + E_{pen}$
3. PARAMETRIZED BY TORSION ANGLES (DIHEDRAL)

RESTRICTED TO THE RANGE -180° to $+180^\circ$
" DECIMAL STRINGS "

4. SELECTION $\left(\begin{array}{l} \text{START} \sim 80\% \\ \text{END} \sim 20\% \end{array} \right)$ Now 100 %
5. MUTATE $10^\circ, 5^\circ, 2^\circ$

START	20 %	10° dominates
END	70 %	2° dominates
6. Crossover

uniform	START	10 %	Uniform
two point	END	70 %	two point

7. FOR AN OCTAPEPTIDE

ALA - THR - TRP - VAL - ASP - LEU - VAL - LEU

8. EVOLVED OVER 1000 GENERATIONS

STARTING FROM RANDOMLY SELECTED ϕ, ψ .

Genetic Algorithms

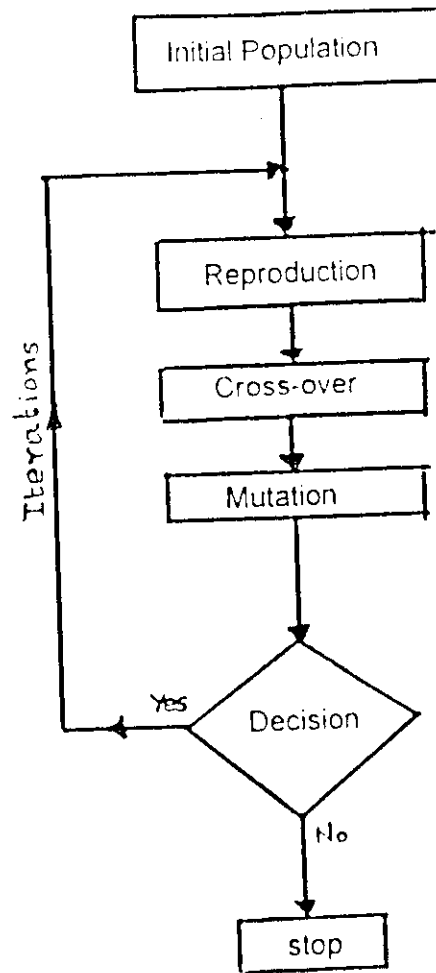
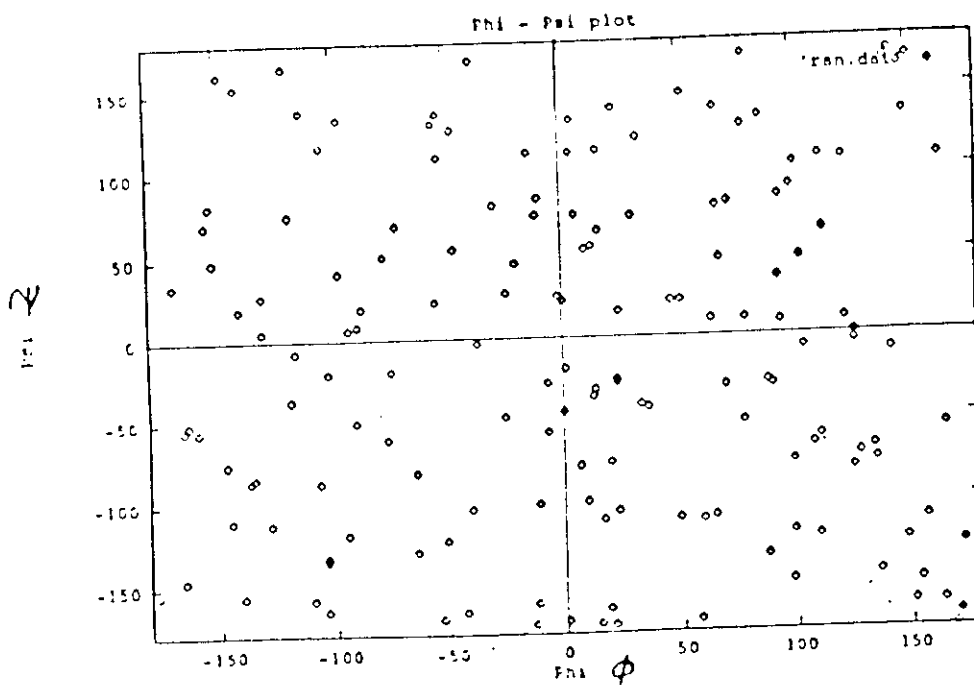


Figure 1



16 Figure 2

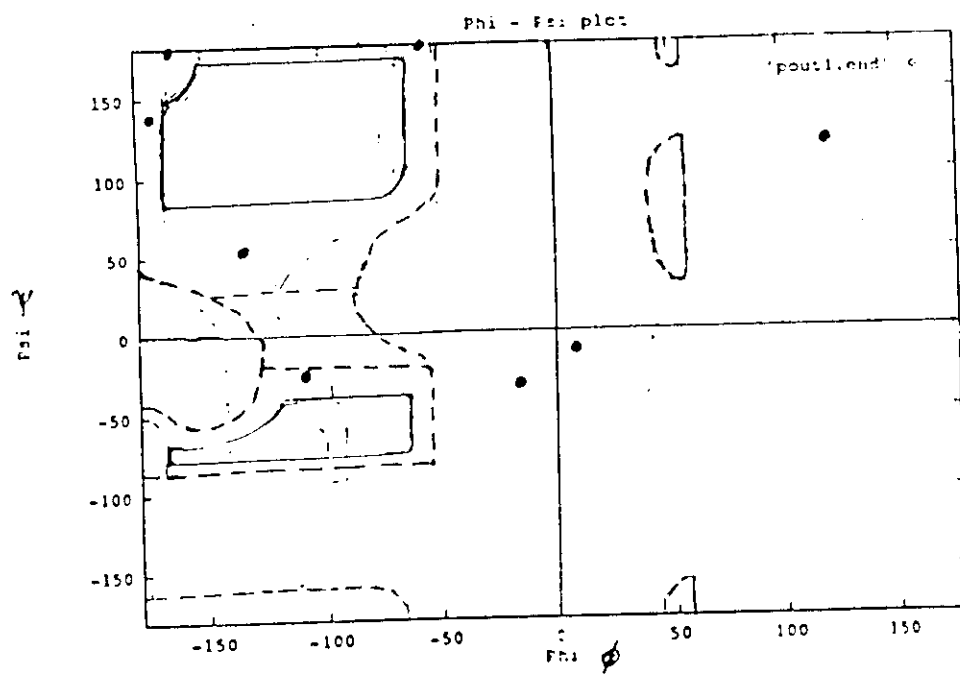


Figure 3

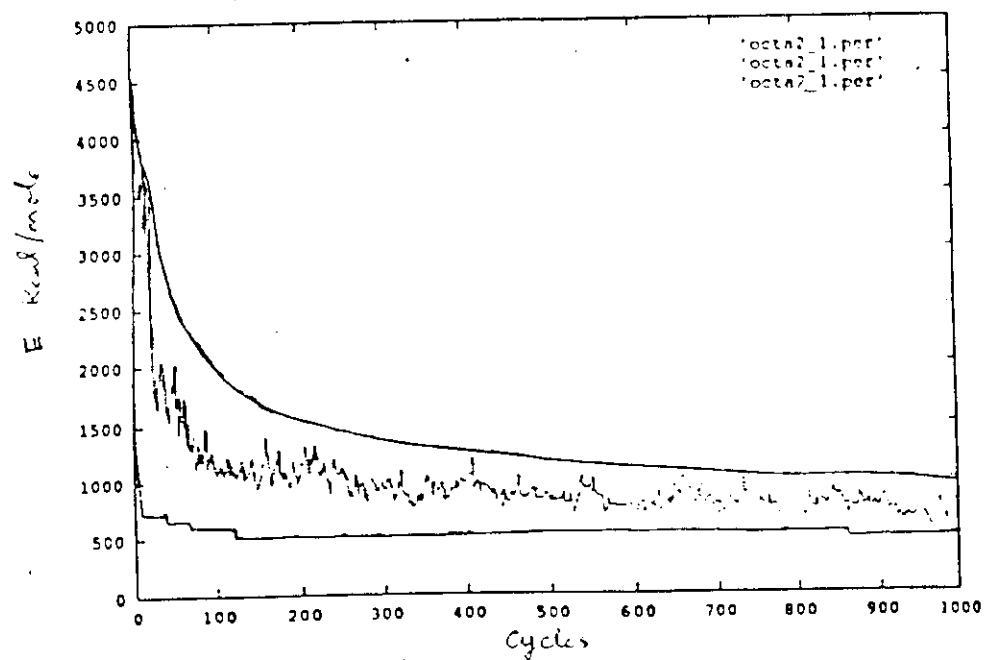


Figure 4

REFERENCES:

- J. H. HOLLAND Adaptation in Natural and Artificial Systems.
Univ. of Michigan Press (1975).
- D. E. GOLDBERG Genetic Algorithms in search, optimisation and Machine learning
NY Addison-Wesley (1989)
- Schulz-Kremer in 'Parallel Genetic Algorithms' Ed by J. Stender
IOS Press (1993)
- SHAOJIAN SUN Protein Science vol 32,
p 762 (1993)
- Dandekar and Patrick Argos
Protein Engg. vol 5 (7) p 637 (92)
- David T. Jones Protein Sci. vol 3, p 567 (94)

