REFINED GENETIC ALGORITHMS
FOR
POLYPEPTIDE STRUCTURE PREDICTION
REFINED GENETIC ALGORITHMS
FOR
POLYPEPTIDE STRUCTURE PREDICTION
Acknowledgements
List of Figures

\begin{center}
\begin{tabular}{ll}
\textit{n} & \text{.....................................} \\
\end{tabular}
\end{center}
List of Tables

\[ \phi, \psi \]

\[ \times 3^0 \]
List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>( \alpha )</td>
</tr>
<tr>
<td>( \mathcal{F} )</td>
<td>( \mathcal{F} )</td>
</tr>
<tr>
<td>( S )</td>
<td>( S )</td>
</tr>
<tr>
<td>( a )</td>
<td>( a )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>( \gamma )</td>
</tr>
<tr>
<td>( i )</td>
<td>( i )</td>
</tr>
<tr>
<td>( \phi )</td>
<td>( \phi )</td>
</tr>
<tr>
<td>( \psi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>( \omega )</td>
<td>( \omega )</td>
</tr>
<tr>
<td>( \chi )</td>
<td>( \chi )</td>
</tr>
<tr>
<td>( i )</td>
<td>( i )</td>
</tr>
</tbody>
</table>
List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstract
REFINED GENETIC ALGORITHMS
FOR
POLYPEPTIDE STRUCTURE PREDICTION

I. Introduction

1.1 Protein Folding Problem / Polypeptide Structure Prediction

predict

native conformation

\[ PSP \subseteq PFP \]

\(^1\)The *native conformation* determines the protein's biological functions.
1.1.1 Background.

primary structure

secondary structure

tertiary structure

1.1.2 Importance.

Grand Challenges
1.1.3 Methods for Polypeptide Structures Prediction.

- energy minimization
- molecular dynamics
- ab initio
- semi-empirical
- force-field

$O(n^3)$ $O(n^4)$ $O(n^9)$

$n$
1.1.4 Growth of Complexity.

\[ x, y, z \]

\[ n \]

\[ [n_0, n] \]

1.2 Genetic Algorithms

*implicit parallelism*\(^3\)

Evolutionary Computing

\(^2\)Regardless of whether a Cartesian or Internal coordinate system is used. However, the internal coordinate system has fewer independent variables.

\(^3\)See Appendix B.3.3

\(^4\)Because GAs are loosely based on natural evolution, many of the terms associated with natural evolution are used interchangeably with the terms created specifically for genetic algorithms [67].
<select, crossover, mutation>

\textit{generation} \hspace{2cm} \textit{genes}

\textit{Alleles}

YES

1.3 Parallel and Distributed Computing

\textit{single instruction stream, multiple data stream}

\textit{multiple instruction stream, multiple data stream}

\textit{scalable}

1.4 Research Objectives

\footnote{\textit{Crossover} is sometimes called \textit{recombination}.}
Effective and Efficient
Polypeptide Structure Prediction

- Improve Performance of Hybrid GAs for PSP
- Real Valued Genetic Algorithm Implementation for the PSP
- Exploit Domain Knowledge to Limit Search Space
  domain knowledge
1.5 Methodology

Program

Evolution

 exogenous

1.6 Assumptions

1.7 Summary

---

6 Not to be confused with Evolutionary Programming, see Section 2.4.
7 The probability that an improvement at a specific node is migrated to other nodes.
8 The probability, given a migration, that it is migrated to all other nodes.
9 Or molecular conformation
II. Current Issues

2.1 Introduction

2.2 Previous Research

2.3 Polypeptides Structure Prediction (PSP)
Permutation Genetic Algorithm

Parallel Implementation (Sawyer)

Communication Strategies (Merkle)

Protein Folding Problem Initial Energy Model (Brinkman)

Protein Folding Problem Energy Model Refinement (Gates)

Messy Genetic Algorithm

Initial Implementation GA-hard Problems (Dymek)

Protein Folding Problem Serial Hybrid GA (Gaulke)

Premature Convergence Strategies (Dymek)

Fast Messy Genetic Algorithm

Combinatoric Optimization Data Distribution Strategies (Merkle)

Additional Parallelization Protein Folding Problem (Gates)

Permutation Genetic Algorithm
\[ \omega \quad \text{trans}^1 \quad \phi, \psi \]

\[ \phi, \psi, \omega \]

\[ \chi^s \]

1 ± 180° or ± π
Ramachandran Plot

ψ

ϕ

Biochemistry
2.4 Genetic Algorithms

program

paradox of real codings

virtual alphabets

blocked

Fundamental Theorem of Genetic Algorithms

strong weak

\[
\begin{array}{ccc}
EP_5 & EP_1 & EP_5 \\
\end{array}
\]

p
2.5 Parallel Genetic Algorithms
data parallelism

island model

neighborhood model

farming model
*Island Model:*

\[ \text{migration} \]

\[ \mathcal{O} \frac{n l}{p} \quad p \quad n \]

\[ p \ll n \quad p \rightarrow n \]

*Neighborhood Model:*

\[ n \quad p \]

\[ \mathcal{O} s \quad \mathcal{O} n l \quad s \]

*Farming Model:*

\[ \text{farming} \]

\[ \text{farm out} \]

\[ \text{foreman} \quad \text{workers} \]

2.6 Summary
III. Algorithm Analysis, Design, and Implementation

3.1 Analysis

3.1.1 Cost Analysis of Local Minimization using Conjugate Gradient.

Numerical Recipes in C

\[ O(n^2) \]

\[ \text{ITMAX} = 200 \]

Lamarckian Baldwin

eval_func()

\[ \text{ITMAX} \]

generational

checkpoint asynchronous

farming
3.1.2 Constraint Set Development.

3.1.2.1 Conventions Adopted.

<table>
<thead>
<tr>
<th>( \theta_{\text{max}} )</th>
<th>( \theta_{\text{min}} )</th>
<th>( \theta_{\text{max}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(- )</td>
<td>(- )</td>
<td>(- )</td>
</tr>
</tbody>
</table>

\( \theta_{\text{max}} \)

\( \theta_{\text{max}} \)

- \( \theta_{\text{min}} > \theta_{\text{max}} \rightarrow \theta_{\min_{a_{\text{d}}}} \)
- \( \theta_{\text{min}} < \theta_{\text{max}} \rightarrow \theta_{\max_{a_{\text{d}}}} \)

\( \theta_{\min_{a_{\text{d}}}} \)

\( \theta_{\max_{a_{\text{d}}}} \)

\( \theta_{\text{min}} \)

\( \theta_{\text{max}} \)

3.1.2.2 [Met]-enkephalin.

\( \phi \psi \)

3.1.2.3 Polysialine.

\( \alpha \)

- \( \alpha \)

\( \phi \psi \)

\( \phi \psi \)

\( \alpha \)
\[
\begin{array}{|c|c|c|}
\hline
 & \theta_{\text{min}} & \theta_{\text{max}} \\
\hline
\text{Non-glycine} & - & - \\
\text{Glycine} & - & - \\
\hline
\chi 1 & - & - \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
 & \theta_{\text{min}} & \theta_{\text{max}} \\
\hline
\text{Non-glycine} & - & - \\
\text{Glycine} & - & - \\
\hline
\chi 1 & - & - \\
\hline
\end{array}
\]

\[
\begin{array}{|c|c|c|}
\hline
 & \theta_{\text{min}} & \theta_{\text{max}} \\
\hline
\text{Non-glycine} & - & - \\
\text{Glycine} & - & - \\
\hline
\chi 1 & - & - \\
\hline
\end{array}
\]

3.1.3 Real-valued GAs.
3.2 Algorithm Design and Implementation

3.2.1 Parallel Hybrid GA.

3.2.1.1 Algorithm Design.

The developers of this algorithm have been inconsistent with its naming in the literature. Genocop, Genocop-III, GENOCOP-III, etc., have been used synonymously. In this document I have adopted the standardization of GENOCOP-III.
\[ Q \]
\[ \text{EvalCnt} \rightarrow \text{population size} \]
\[ \text{Loop} \]
\[ \text{Until} \text{EvalCnt} \]
\[ \text{Loop} \]

\[ \text{EvalCnt} \rightarrow \]
\[ \text{Loop} \]
\[ \text{Until} \text{EvalCnt} \]
\[ \text{Until} \]

3.2.1.2 Scheduling.

\text{round robin}

\text{network of workstations}

3.2.2 \text{REal-valued GA, Limited by constraints (REGAL)}.
Loop

Until

\[ \text{REPLACE} \]

\[ \text{serve}r_i \quad Q \]

\[ \text{EvalsPerfomed} \]

\[ \text{EvalsPerfomed} < \text{EvalCnt} \land Q \not\subseteq \emptyset \]

\[ \text{serve}r_j \quad Q \]

\[ \text{serve}r_j \]

\[ \text{3.2.2.1 Incorporation of Domain Knowledge.} \]

\[ \text{Evolution Program} \]

\[ \text{probable} \quad \text{improbable} \quad S_{prob} \]

\[ S_{improb} \]

\[ S_{prob} \quad S_{improb} \quad S \]

\[ S_{prob} \quad S_{improb} \quad \emptyset \]
\[ x_i \in [\theta_{\text{min}}, \theta_{\text{max}}] \quad \text{for} \quad x_i \in \mathbb{R} \]

\[ \theta \leq \theta_{\text{min}}, \theta_{\text{max}} - \theta_{\text{max}} - \theta_{\text{min}} \]

\[ \{ \phi, \psi, \omega \} \]

\[ \leq \theta - \theta_{\text{min}}, \theta_{\text{max}} - \theta_{\text{max}} - \theta_{\text{min}} \]

\[ \{ x_1 \} \]

3.2.3 Parallel REGAL (Para-REGAL).

modified island

island

\[ P_m \] Probability of Complete Migration \( P_{cm} \)

\[ P_m \cdot P_{cm} \]

\[ P_m \cdot P_{cm} > . \]

archipelago \(^3\)

\begin{itemize}
  \item \textbf{Average Genotypic Distance}
  \item \textbf{Least Genotypic Distance}
\end{itemize}

\(^2\)Except for the seed of the random number generator.

\(^3\)A collection of islands; thus the islands making of the Para-REGAL execution
- Greatest Genotypic Distance

- Average Best Fitness

- Local Delta

3.3 Summary
IV. Experiment Design

4.1 Experiment Techniques

4.1.1 Random Number Seeds.

nodal

Seed = (unsigned) ((Seed + My_node) / (My_node + 1));
\[ \text{Seed} = (\text{unsigned}) \left( \left( \text{Seed} + \text{My}_\text{node} \right) \% \text{Max}_\text{Seed}_\text{Value} \right); \]

\[ \text{Seed} \quad \text{floor} \quad * \text{rand number from } X\text{calculator} \]

\[ \text{Seed} \quad \text{floor} \quad * \text{random number from } X\text{calculator} \]

\[ \text{Seed} \quad \text{floor} \quad * \text{random number from } X\text{calculator} \]

4.1.2 Statistical Techniques.

4.1.2.1 Analysis of Variance (ANOVA). 

\[ n \quad a \]

\[ a \]

\[ \sigma^2 \]
4.1.2.2 Kruskal-Wallis H Test.

4.2 Experiment I: Evaluation of the Efficiency of a Parallel & Distributed Hybrid GA

4.2.1 Motivation and Objective.

4.2.2 Methodology.

\[ P_m \]

\[ P_r \]

Farmer model
4.3 Experiment II: Evaluation of the Use of Constraints in the PSP

4.3.1 Motivation and Objective.

4.3.2 Methodology.

\[ \text{none} \quad \text{loose} \quad \text{tight} \]

\[ [\text{Met}]^\text{-enkephalin} \quad \text{tight} \]

\[ \text{Polyalanine} \quad \text{tight} \]

4.3.3 Parameter Selection.

\[ \text{reference population} \]

\[ \mathcal{F} \quad \mathcal{S} \]
<table>
<thead>
<tr>
<th>loose</th>
<th>tight</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4.4 Experiment III: Evaluation of Exogenous Parameters in the REGAL System

4.4.1 Motivation and Objective.

\[^1\]Based on the ratio of $\frac{F}{S}$, it would require $10^{17}$ tries to randomly generate just on fully feasible chromosome when using the tight constraint set for Polyalanine.
4.4.2 Methodology.

*steady state*

*adaptive*

*iterations*
### 4.4.3 Exogenous Parameter Evaluation Experiments.

<table>
<thead>
<tr>
<th>Offsprings</th>
<th>Reference Population Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4.5 Experiment IV: Evaluation of Para-REGAL

#### 4.5.1 Motivation and Objective.
4.5.2 Methodology.

| Island 0 |
| Islands 1 & 2 | loose | Island 3 | tight |

4.5.3 Para-REGAL Experiments.

\{ \ldots, \ldots, \ldots \} \quad P_m \quad P_{cm} 

\omega
V. Results and Analysis

5.1 Experiment I: Parallel Hybrid GA

5.1.1 Effectiveness Analysis.

\[ \frac{x^{\text{pop size small}}}{x^{\text{pop size large}}} > \frac{x^{\text{pop size large}}}{x^{\text{pop size small}}} \]

\[ \text{selection pressure} \]
5.1.2 Efficiency Analysis, Serial.
5.1.3 Efficiency Analysis, Parallel.

\[
C_{\text{communications}} \quad T_P - T_S
\]

\[
\frac{T_P}{T_S}
\]

Communication Cost Index \[
\frac{T_P}{T_S}
\]

![Graph showing communication cost index over evaluations]

\[
S \quad \frac{T_S}{T_P}
\]
\[ S \quad T_S \quad \text{best} \quad T_p \]

linear speedup \( S \)

\[ p \quad S > p \quad \text{super linear speedup} \quad T_S \]

---

**Graphs:**

- **Graph 1:** Speedup vs. Number of Processors for different algorithms (FPBald, FPLam, TSLam, FPSGA, TSSGA).
- **Graph 2:** Speedup vs. Number of Processors for a different set of algorithms, showing varied performance trends.
5.2 Experiment II: Preliminary REGAL Evaluation
5.2.1 \([\text{Met}]\)-enkephalin.
5.2.2 Polyalanine.

\[ \alpha \]

\[ \alpha \]

5.2.3 Efficiency.
5.3 Experiment III: Analysis of Exogenous Parameters for REGAL

independent

current best trajectory

\(^2\)Analysis of Variance, see Appendix F.1 for more details. Concern has been raised about lack of variability because a single
seed set was used. The Kruskal-Wallis H Test (Appendix F.2) was used as an independent method to verify the ANOVA results.
The conclusions were the same. Kruskal-Wallis results are not shown.

\(^3\)Hypothesis testing was not done on run times because system loading in the multi-user environment could not be controlled.
They are provided for reference only. However, the large number of experiments tends to dampen out cases were the platform
was heavily loaded. Thus, the data are insightful.
This nomenclature is from the GENOCOP-III documentation. It would be more accurate to say the reference population is operated upon.
\[
\begin{array}{cccc}
F_{\parallel} & a & \cdots & \\
\hline
\hline
\end{array}
\]

\[
\begin{array}{cccc}
\multicolumn{3}{c}{H_{\parallel}} & \ast \\
\hline
\hline
\end{array}
\]

\[
\begin{array}{cccc}
a & \cdots & a & \ast \\
\hline
\hline
\end{array}
\]

\[
\begin{array}{cccc}
a & \cdots & a & a^{-i} \leq i \leq \ast \\
\hline
\hline
\end{array}
\]
5.4 Experiment IV: Analysis of Para-REGAL

\[ P_m \quad P_{m_n} \quad \alpha \]

Island 3
periodicity
\[
\begin{align*}
    n & \quad \frac{m}{n} & \quad m \\
    -3 & \quad th & \\
    - & \quad K
\end{align*}
\]

<table>
<thead>
<tr>
<th>( P_m )</th>
<th>( P_{cm} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.5 Summary

Considerable resources is a relative concept. Even 1000 hours of computer time is trivial when compared to experimental techniques that require years to yield results.
VI. Conclusions and Recommendation

Grand Challenge

6.1 Initiative I: PHGA

6.2 Initiative II: REGAL

\[ \text{trans} \]

\[ \text{trans} \]

---

1 Actual implementation is out of scope for this investigation because research is required into appropriate control metrics.
2 Limit the dihedral angle's range to a lower bound greater than \(-\pi\) and an upper bound less than \(\pi\).
fundamental theorem of genetic algorithms

6.3 Initiative III: Examination of Exogenous Parameters

6.4 Initiative IV: Para-REGAL

molten globular
Genetic Fast Messy Algorithm
Initial Implementation
Parallel Implementation (Merkle)
Additional Parallelization Protein Folding Problem (Gates)
Protein Folding Problem Initial Energy Model (Brinkman)
Initial Implementation GA-hard Problems (Dymek)
Mathematical Model (Merkle)
Exogenous Parameter Selection (Merkle)
Linkage Friendly Genetic Algorithms
Parallel Implementation (Sawyer)
6.5 Recommendations

6.6 Summary
Appendix A. Background on the Protein Folding and Protein Structure Prediction Problems

inverse folding

problem

evaluation

design

A.1 Introduction to Proteins and Associated Terminology

backbone

side-chain $i$

primary structure

residues
### Secondary Structure

| secondary structure | $\phi$ | $\psi$
|---------------------|--------|--------
| $\alpha$            |        |        |
| $\beta$             |        |        |
| $\gamma$            |        |        |
| $\delta$            |        |        |
| $\epsilon$          |        |        |
| $\zeta$             |        |        |

### Tertiary Structure

<table>
<thead>
<tr>
<th>tertiary structure</th>
<th>conformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$\epsilon$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>$\zeta$</td>
</tr>
</tbody>
</table>

### Table of Bond Angles and Bond Lengths

<table>
<thead>
<tr>
<th>Bond Angle</th>
<th>Bond Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>$\gamma$</td>
</tr>
<tr>
<td>$\phi$</td>
<td>$\delta$</td>
</tr>
<tr>
<td>$\psi$</td>
<td>$\epsilon$</td>
</tr>
</tbody>
</table>

### Nomenclature

- **C**: Carbon
- **H**: Hydrogen
- **N**: Nitrogen
- **O**: Oxygen
native

\[ x_i, y_i, z_i \quad \leq i \leq n \quad n \]

\[ C_{\alpha_i} \]

\[ d_{ij} = \frac{x_i - x_j}{y_i - y_j} \frac{z_i - z_j}{n} \]

bond length

bond angle

dihedral angle

\[ \phi \quad \psi \quad \omega \quad \chi_i \]

n - \quad n
A.2 Experimental Tertiary Structure Determination

A.3 Tertiary Structure Prediction (PFP)
A.3.1 Classical Prediction Methods. Molecular dynamics

extended-atom representation

energy minimization

<table>
<thead>
<tr>
<th></th>
<th>( O )  ( n^5 )</th>
<th>( O )  ( n^4 )</th>
<th>( O )  ( n^3 )</th>
<th>( O )  ( n^2 )</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ab \ initio )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n )</td>
</tr>
<tr>
<td>semi-empirical</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n )</td>
</tr>
<tr>
<td>force-field</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>( n )</td>
</tr>
</tbody>
</table>
\[ E \frac{U_{ijkl}}{r_{ij}} \pm \frac{n_{ijkl}}{i j k l} \]

\[ \epsilon_{ij} \frac{F_{ij}}{r_{ij}} \frac{r_0}{12} - \frac{r_0}{6} \]

\[ \frac{q_i q_j}{D r_{ij}} \]

\[ \frac{r_0}{r_{ij}} \frac{r_0}{r_{ij}} \]

\[ \frac{r_0}{r_{ij}} \frac{r_0}{r_{ij}} \]

\begin{itemize}
  \item \( \mathcal{D} \)
  \item \( \mathcal{N} \)
  \item \( \mathcal{H} \)
  \item \( r_{HX} \)
  \item \( r_{ij} \)
  \item \( i j k l \)
  \item \( q_i \)
  \item \( U_{ijkl} \)
\end{itemize}

\[ E_B \ E_A \ E_D \ E_N \ E_N' \]

A.3.2 Other Prediction Methods.

homology

sequence-structure alignment
\[
E = \sum_{(i,j) \in B} K_{r_{ij}} r_{ij} - r_{eq}^2
\]
\[
+ \sum_{(i,j,k) \in \mathcal{A}} K_{\Theta_{ijk}} r_{ijk} - \varepsilon_{eq}^2
\]
\[
+ \sum_{(i,j,k,l) \in \mathcal{D}} K_{\Phi_{ijkl}} r_{ijkl} - \varepsilon_{ijkl}^2
\]
\[
- \sum_{(i,j) \in \mathcal{N}} \frac{A_{ij}}{r_{ij}} \left( \frac{B_{ij}}{r_{ij}} - \frac{q_i q_j}{\pi \varepsilon_{ij}} \right)
\]
\[
- \sum_{(i,j) \in \mathcal{N}'} \frac{A_{ij}}{r_{ij}} \left( \frac{B_{ij}}{r_{ij}} - \frac{q_i q_j}{\pi \varepsilon_{ij}} \right)
\]

- $B$
- $\mathcal{A}$
- $\mathcal{D}$
- $\mathcal{N}$
- $\mathcal{N}'$
- $r_{ij}$
- $i, j$
- $ijk$
- $i, j, k$
- $ijkl$
- $i, j, k, l$
- $q_i$
- $i$
- $K_{r_{ij}}$
- $r_{eq}$
- $K_{\Theta_{ijk}}$
- $\varepsilon_{eq}$
- $K_{\Phi_{ijkl}}$
- $\varepsilon_{ijkl}$
- $A_{ij}$
- $B_{ij}$
- $q_i q_j$
- $\pi \varepsilon_{ij}$
Simplification

Lattice
Appendix B. Background on Genetic Algorithms

population encoded operators fitness
generation

always

irregular

good
B.1 Brief History of Evolutionary Algorithms

Evolutionstrategie

\[ V \langle x, \sigma \rangle \quad x \quad \sigma \quad t \]
\[ x^{t+1} = x^t + N \cdot \sigma \]

\[ \sigma \]

B.2 Origins of Genetic Algorithms

"Adaptation in Natural and Artificial Systems"

Schema Theorem  Fundamental Theorem of Genetic Algorithms

B.3 Simple Genetic Algorithm (SGA)

\[ \text{string} \quad \text{chromosome} \quad \text{genes} \]

\[ \text{locus} \quad \text{allele} \]

\(^1\text{The term phenotype refers to the traits expressed by an individual, in this case the value returned by a function. Contrast this with genotype which refers to the traits that define the individual, for example the parameters of the function.}\)
Population evolves

### Population

<table>
<thead>
<tr>
<th>Locus (Position)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome (String)</td>
<td>101111100</td>
<td>110011110</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome (String)</td>
<td>000101011</td>
<td>001101011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome (String)</td>
<td>010010111</td>
<td>011101100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chromosome (String)</td>
<td>011001001</td>
<td>101100010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Simple Genetic Algorithm Operators

- **Selection**
- **Crossover**
- **Mutation**

### Formulas

- $x_{min}$
- $x_{max}$

### B.3.1 Simple Genetic Algorithm Operators

- **Selection**
- **Crossover**
- **Mutation**

**Converge**
**single-point crossover**  

**bitwise mutation**

---

**CROSSOVER POINT**

**PARENT #1**  
0 0 1 0 1 1 1 0 1 1  

**PARENT #2**  
0 1 1 1 0 1 0 1 1 0  

**CHILD #1**  
0 0 1 0 1 1 0 1 1 0  

**CHILD #2**  
0 1 1 1 0 1 1 0 1 1  

---

**MUTATION POINT**

**INPUT**  
0 0 1 0 1 1 1 0 1 1  

**OUTPUT**  
0 0 1 0 1 0 1 0 1 1  

---

**proportional**  

**roulette-wheel**

---

<table>
<thead>
<tr>
<th>String</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>12</td>
</tr>
<tr>
<td>S2</td>
<td>12</td>
</tr>
<tr>
<td>S3</td>
<td>12</td>
</tr>
<tr>
<td>S4</td>
<td>12</td>
</tr>
<tr>
<td>mean</td>
<td>12</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>String</th>
<th>Fitness</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>20</td>
</tr>
<tr>
<td>S2</td>
<td>10</td>
</tr>
<tr>
<td>S3</td>
<td>5</td>
</tr>
<tr>
<td>S4</td>
<td>5</td>
</tr>
<tr>
<td>mean</td>
<td>10</td>
</tr>
</tbody>
</table>
Rank-based tournament elitist

Recombination
Mutation
Selection
B.3.2 Simple Genetic Algorithm Parameters.

B.3.3 Mathematical Theory of How (Why) Simple GAs Work. Schemata
don't care

{ } defining length \( \delta H \) order

\( o H \)

\( H \)

\( \delta * \) \( - \) \( o * \) \( \delta * * \) \( - \)

\( o * * \)
\[
m H, t \geq m H, t \cdot \frac{f H}{f} - p_c \frac{\delta H}{t} - o H p_m, \quad H
\]

\[
m H, t \quad \frac{f((H))}{f} \quad H \quad p_c \frac{\delta|H|}{t}
\]

\[
H \quad o H p_m \quad i \quad implicit
\]

parallelism

B.3.3.1 Complexity Analysis.

\[
O \text{ nl} \quad n \quad l \quad O \text{ nl}
\]

\[
O \text{ nl}
\]
B.4 Messy Genetic Algorithm (mGA)

**building block**

**Deception**

**linkage**

*under-specified*  *over-specified*

**competitive template**

B.4.1 Messy Genetic Algorithm Operators.

**thresholding**

*cut-and-splice*

*cut*  *splice*

**primordial**  *juxtapositional*  *partially enumerative initialization*  *(PEI)*

\[ k \sim i \quad k \]

B.4.2 Messy Genetic Algorithm Parameters.

normalized expected defining length $\frac{\langle \ell \rangle}{\ell + 1} k$

\[
\frac{\langle \ell \rangle}{l} = \frac{k - 1}{k}
\]

B.4.3.1 Complexity Analysis.

$O^{1k}$

$O \mid l \mid l$

$n \quad \text{much}$

B.5 Fast Messy Genetic Algorithm (fmGA)

B.5.1 Fast Messy Genetic Algorithm Operators.

probabilistically complete initialization
B.5.2 Fast Messy Genetic Algorithm Parameters.

B.5.3.1 Complexity Analysis.

\[ \mathcal{O}(l^k) \]
Appendix C. Background on Parallel Computing

C.1 Parallel Architectures
shared memory

distributed memory

message passing

interconnection topology network

2-D mesh

\times

hypercube

dimension \ N

N

N

N
C.2 Parallel Algorithms.
in in.something
something
in

in.something

> 
/dev/null

psga_default  psga_param
psga_default
in
mpirun

-np  -sz

in.2.10.24
&  >
/dev/null
Experiments = 1
Total Trials = 500
Population Size = 20
Structure Length = 240
Crossover Rate = 0.65
Mutation Rate = 0.005
Generation Gap = 1.0
Scaling Window = 1
Report Interval = 1
Structures Saved = 1
Max Gens w/o Eval = 10
Dump Interval = 0
Dumps Saved = 0
Options = ycel
Number of Peaks = 1.0
Minimization Prob = 1.0
Replacement Prob = 1.0
Random Seed = 987654321
Rank Min = 1.5
E.1 Algorithm

Procedure Genocop III

begin
    \( t \leftarrow t \)
    \( P_s t \)
    \( P_r t \)
    \( P_s t \)
    \( P_r t \)

while not do
    begin
        \( t \leftarrow t \)
        \( P_s t \)
        \( P_r t \)
        \( P_s t \)
    end

if \( t \text{ mod } k \) then
    begin
        \( P_r t \)
        \( P_r t \)
        \( P_r t \)
    end
end
end
procedure evaluate $P_s t$
begin
\[ \hat{s} \in P_s t \] do
\[ \text{if } \hat{s} \in F \] then
\[ \bar{s} \\ f \hat{s} \] else
begin
\[ \bar{r} \in P_r t \]
\[ \bar{z} \in F \]
\[ \hat{s} \\ f \bar{z} \]
\[ \text{if } f \bar{r} > f \bar{z} \] then
\[ \bar{r} \quad \bar{z} \quad P_r \]
\[ \hat{s} \quad \bar{z} \quad P_s \quad p_r \]
end
end
<table>
<thead>
<tr>
<th></th>
<th>( N \cup )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( N \cup )</td>
</tr>
<tr>
<td></td>
<td>( N \cup )</td>
</tr>
<tr>
<td></td>
<td>( N \cup )</td>
</tr>
<tr>
<td></td>
<td>( N \cup )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td></td>
<td>( N )</td>
</tr>
<tr>
<td>( N \cup )</td>
<td>( N \cup )</td>
</tr>
</tbody>
</table>

**E.3 Operators**
E.3.1 Whole arithmetical crossover.
\[ z = a\bar{x} - a\bar{y} \leq a \leq \]
\[ \bar{x} \quad \bar{y} \]

E.3.2 Simple arithmetical crossover.

\[ \text{single point crossover} \]

E.3.3 Whole uniform mutation.
\[ \bar{x} \quad x_1, \ldots, x_k, \ldots x_n \]
\[ \bar{x}' \quad x_1, \ldots, x_k, \ldots x_n \]
\[ k \]

E.3.4 Boundary mutation.
\[ x_k \quad \text{left } k \quad \text{right } k \]

E.3.5 Non-uniform mutation.
\[ x_{k+1}^i \quad \Delta t, r \quad x_k \quad \Delta t, k \]
\[ x_k^i - \Delta t, x_k - l \quad k \]
\[ \Delta t, y \quad \ldots, n \quad \Delta t, y \]
\[ t \quad \triangle t \quad t \]
\[ t \]
\[ \Delta t, y \quad y \cdot r \cdot - \frac{t}{T} b \]
\[ r \quad \ldots \quad T \quad b \]
E.3.6 Whole non-uniform mutation.

E.3.7 Heuristic crossover.

\[ \hat{x} \quad \bar{y} \]
\[ \hat{z} = r \cdot \hat{x} - \bar{y} \quad \hat{x} \]
\[ r \]
\[ f \hat{x} \leq f \bar{y} \]

E.3.8 Gaussian mutation.

\[ \bar{x} \quad \langle x_1, \ldots, x_n \rangle \]
\[ \bar{x}^{t+1} = \bar{x}^t + N, \sigma \]
\[ N, \sigma \]
\[ \bar{\sigma} \]

E.3.9 Pool recombination operator.

E.3.10 Scatter search operator.

\[ k > J \]
\[ \bar{b} \quad \bar{w} \]
\[ \bar{c} \]
\[ \bar{c} \]
\[ \bar{c}_{i=J-w} \]
\[ \bar{y} \]
\[ \bar{y} \quad \bar{c} \quad \bar{c} - \bar{w} \]
Appendix F. Statistical Methods

F.1 Analysis of Variance (ANOVA)

F.1.1 Single Factor Factorial Design.

\[ y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad i, \ldots, a; \quad j, \ldots, n \]

\[ y_{ij} = \mu + \epsilon_{ij}, \quad i \quad \text{treatment effect} \quad \epsilon_{ij} \]

\[ \sigma^2 \]

\text{effects model only}

\text{fixed effects model}

\text{random samples}

\text{random effects model} \quad \tau_i \]

\[ \tau_i \]

\[ \alpha \quad n \]

\[ N \]

\text{analysis of variance}

\[ SS_T = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 \]

\[ SS_T \]
\[
SS_T \quad SS_{Treatments} \quad SE
\]

\[
SS_{Treatments} \quad SE
\]

\[
H_0 \quad \tau_1, \tau_2, \ldots, \tau_a
\]

\[
F_0 \quad \frac{SS_{Treatment} / a - \left(\bar{y}^2 - \frac{y_{ij}^2}{N}\right)}{SS_E / N - a} \quad \frac{MS_{Treatments}}{MS_E}
\]

\[
F \quad a - N - a
\]

\[
H_0
\]

\[
F_0 > F_{a-1, N-a}
\]

\[
SS_{Treatments} \quad \sum_{i=1}^{a} \sum_{j=1}^{n} \frac{y_{ij}^2}{N}
\]

\[
SS_E \quad \sum_{i=1}^{a} \frac{\bar{y}_{ii}^2}{n} - \frac{\bar{y}^2}{N}
\]

\[
SS_T \quad SS_{Treatments} - SS_{Treatments}
\]

\[
F_0
\]

\[
SS_{Treatments} \quad a - \quad MS_{Treatments} \quad F_0 \quad MS_{Treatments} \quad MS_E
\]

\[
SS_E \quad N - a \quad MS_E
\]

\[
SS_T \quad N -
\]

\[
y_i \quad \sum_{j=1}^{n} y_{ij}, \quad \bar{y}_i, \quad \bar{y}_i / n \quad i, \ldots, a
\]
F.1.2 Two Factor Factorial Design.

\[
\begin{align*}
  &y_{ij} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \epsilon_{ijk} \\
  &\quad i = 1, \ldots, a \\
  &\quad j = 1, \ldots, b \\
  &\quad k = 1, \ldots, n \\
  &\mu, \tau_i, \beta_j, \tau\beta_{ij}, \epsilon_{ijk} \\
  &\quad A, \beta_j \\
  &\quad B, \tau\beta_{ij} \\
  &\quad A \\
  &H_0: \tau_1 = \tau_2 = \ldots = \tau_a \\
  &H_1: \exists \tau_i / \\
  &\quad B \\
  &H_0: \beta_1 = \beta_2 = \ldots = \beta_b \\
  &H_1: \exists \beta_i / \\
  &A, B \\
  &H_0: \tau\beta_{ij} \quad \forall i, j \\
  &H_1: \exists \tau_i / \\
  &H_0
\end{align*}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>Degree of Freedom</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( SS_A )</td>
<td>( a - 1 )</td>
<td>( a - 1 )</td>
<td>( a - 1 )</td>
<td>( a - 1 )</td>
</tr>
<tr>
<td>( SS_B )</td>
<td>( b - 1 )</td>
<td>( b - 1 )</td>
<td>( b - 1 )</td>
<td>( b - 1 )</td>
</tr>
<tr>
<td>( SS_{AB} )</td>
<td>( ab - 1 )</td>
<td>( ab - 1 )</td>
<td>( ab - 1 )</td>
<td>( ab - 1 )</td>
</tr>
<tr>
<td>( SS_E )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
</tr>
<tr>
<td>( SS_T )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
<td>( abn - 1 )</td>
</tr>
</tbody>
</table>
\[ SS_T = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk}^2 \]

\[ SS_A = \frac{1}{bn} \sum_{i=1}^{a} \frac{1}{n} \sum_{j=1}^{b} y_{ij.}^2 \]

\[ SS_B = \frac{1}{an} \sum_{j=1}^{b} \frac{1}{n} \sum_{i=1}^{a} y_{i.j}^2 \]

\[ SS_{AB} = SS_{Subtotals} - SS_A - SS_B \]

\[ SS_{E} = SS_T - SS_{Subtotals} \]

\[ SS_{Subtotals} = \frac{1}{n} \sum_{i=1}^{a} \frac{1}{n} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ij.k}^2 \]

\[ y_{.i} = \frac{1}{bn} \sum_{j=1}^{b} y_{ij.} \]

\[ y_{.j} = \frac{1}{an} \sum_{i=1}^{a} y_{i.j} \]

\[ y_{..i} = \frac{1}{an} \sum_{j=1}^{b} y_{ij.} \]

\[ y_{..j} = \frac{1}{bn} \sum_{i=1}^{a} y_{i.j} \]

\[ y_{..} = \frac{1}{abn} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{n} y_{ijk} \]

\[ k \]

\[ n \]

\[ R_{i} \]

\[ i \]
Suppose we have \( k \) independent samples from \( k \) populations. We wish to test the null hypothesis

\[ H_0 : \text{the samples are from identical populations} \]

against the alternative hypothesis

\[ H_1 : \text{the populations are not identical} \]

at the \( \alpha \) level of significance.

1. Compute \( h \). Calculate

\[
h = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)
\]

2. Accept or reject \( H_0 \). If \( h > \chi^2_{k-1, \alpha} \), reject \( H_0 \); otherwise accept \( H_0 \).

(1)
Probability, Statistics, and Queuing Theory: With Computer Science Applications

Journal of Global Optimization

Evolutionary Algorithms in Theory and Practice
Algorithmics: Theory and Practice

Journal of Computational Chemistry

Journal of Global Optimization

Parallel Program Design: A Foundation

Proceedings of the Fourth International Conference on Genetic Algorithms

Physics Today

Proceedings of the Fifth International Conference on Genetic Algorithms

An Analysis of the Behavior of a Class of Genetic Adaptive Systems

Handbook of Genetic Algorithms

IEEE Transactions on Systems, Man and Cybernetics
Genetic Algorithms and
their Applications: Proceedings of the Second International Conference on Genetic Algorithms

Parallel Genetic Algorithms

The Second Annual Conference on Evolutionary Programming

Task Scheduling in Parallel and Distributed Systems

Proceedings of the Third International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

The Mathematica Journal

International Conference on Genetic Algorithms

Proceedings of the Second IEEE Conference on Evolutionary Computation

The Second Annual Conference on Evolutionary Programming

Artificial Intelligence Through Simulated Evolution

IEEE—ACSSC

Foundations of Genetic Algorithms 2

Decision Sciences

*Genetic Algorithms in Search, Optimization, and Machine Learning*

of Populations

*Genetic Algorithms, Noise, and the Sizing*

the *Fifth International Conference on Genetic Algorithms*

*Complex Systems*

Conference on Genetic Algorithms

*Complex Systems*

Proceedings of the Second International Conference on Genetic Algorithms

Proceedings of the *Fifth International Conference on Genetic Algorithms*

*IEEE Transactions on Systems, Man & Cybernetics*

Proceedings of the Third International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

Deception Considered Harmful

*Proceedings of International Parallel Processing Symposium*

Adaptation in Natural and Artificial Systems

*Scientific American*
Principles of Biochemistry

Biochemistry


Applied Computing

The Origins of Order, Self-Organization and Selection in Evolution

Advances in Parallel Algorithms

Design and Analysis of Algorithms

Introduction to Parallel Computing:

Compendium of Parallel Programs for the Intel iPSC Computers

Calculus with Analytic Geometry

Journal of Global Optimization

The Protein Folding Problem and Tertiary Structure Prediction

Arbeitspapiere der

GMD 748

Real Time System Design

Introduction to Parallel Computing
Proceedings of the National Academy of Science USA

Theory and Problems of Organic Chemistry


Evolutionary Computation

Proceedings of the 3rd Annual Conference on Evolutionary Programming

Evolutionary Computation Journal

Genetic Algorithms + Data Structures = Evolution Programs

Genetic Algorithms + Data Structures = Evolution Programs

Molecular Structure (Theochem)

Design and Analysis of Experiments

Parallel Computing

Journal of Computational Chemistry
Proceedings 1

Heuristics

Proceedings of the Third International Conference on Genetic Algorithms

in C: The Art of Scientific Computing

Protein Science

International Conference on Genetic Algorithms

Proceedings of the Fourth International Conference on Genetic Algorithms

Reference

WEBSTER'S II New Riverside University Dictionary

IEEE-CH

Proceedings of the Fourth International Conference on Genetic Algorithms

COMPUTER

Proceedings of the Fourth International Conference on Genetic Algorithms

Biochemistry

genitor
International Conference on Genetic Algorithms

PPSN III

Proceedings of the Fourth International Conference on Genetic Algorithms

Foundation of Genetic Algorithms
Vita