Tutorial on Evolutionary Computation in Bioinformatics: Part I

Gary B. Fogel  
Vice President  
Natural Selection, Inc.  
9330 Scranton Rd., Suite 150  
San Diego, CA 92121 USA

Kay C. Wiese  
School of Computing Science  
Simon Fraser University  
Canada  
Chair: IEEE CIS Bioinformatics and Bioengineering Technical Committee
Bioinformatics - Definition

- Bioinformatics
  - The field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.
  - Classically: storage and information retrieval of biological data

- Computational Biology
  - Use of computers to analyze and interpret biological data
  - Typically nucleotide, RNA, protein sequences, structures

- Largely interchangeable in the literature

- New tools for data access and management

- New algorithms and statistics for pattern identification and prediction
Bioinformatics - History

- Revolutionary methods in molecular biology
  - DNA sequencing
  - Protein structure determination
  - Drug design and development
- Exponential growth of biological information
- Computational requirement for
  - Database storage of information
  - Organization of information
  - Tools for analysis of data
- The transition of biology from “wet-lab” to “dry-lab/information science”
Bioinformatics / Computational Biology

- The use of techniques from applied mathematics, informatics, statistics, and computer science to solve (typically noisy) biological problems

- Multiple sequence alignment
- Identification of functional regions or motifs
- Classification of data
- Phylogenetic analysis
- Molecular structure determination and folding
- Genetics
- Diagnostics and medical applications
Why is bioinformatics important?

- Modeling via bioinformatics may provide answers to questions related to human health and evolution
- Development of small molecule drugs to combat infection and disease
- Diagnosis and prognosis, leading to more effective medicine
Bioinformatics

The Knowledge Gap

Bioinformatics + wet work operate in this gap

Known structures, functions, etc ...

Data

Time

Amount

the 50’s

now
Information Flow in the Cell

Transcription → Translation

DNA → RNA → Protein

Reverse transcription

Information → Function
A More Modern View…
What is DNA?

- Adenine (A)
- Thymine (T)
- Guanine (G)
- Cytosine (C)

- A – T
- G – C
Gene Organization

Enhancers and promoters affect the level of transcription and act as on-off switches (potentiometers)
Promoter Data

- Reese' promoters dataset:
- [http://www.fruitfly.org/seq_tools/datasets/Human/promoter/](http://www.fruitfly.org/seq_tools/datasets/Human/promoter/)
- Results for NNPP on promoters taken from the Eukaryotic Promoter Database, EPD and genes GenBank database.
- 300 promoter sequences of 51 bp each. (40bp upstream and 11 bp downstream from the known transcription start site)
- 3,000 non-promoter regions, also each of 51 bp, some from coding regions and some introns.
Introns

DNA

Exons

Transcription, elimination of intron transcript segments, and splicing of exons

mRNA
RNA

- Adenine (A)
- Uracil (U)
- Guanine (G)
- Cytosine (C)

- A – U
- G – C
- G – U
>100 Gigabases of Information

- Over 100 Gigabases of DNA and RNA sequence information in GenBank, EMBL, and DDBJ as of 2005
  - (roughly the same order of magnitude as the number of nerve cells in a human brain)
  - individual genes
  - partial and complete genomes
  - over 165,000 organisms
  - Free access

Databases

- GenBank
  - National Institutes of Health
  - 65 gigabases of sequence information
- EMBL
  - European Molecular Biology Laboratory
- DDBJ
  - DNA DataBank of Japan
Microarrays
1.28 cm

Actual size of GeneChip™

Millions of DNA strands built up in each cell

500,000 cells on each GeneChip™ array

Actual strand = 25 base pairs
RNA fragments with fluorescent tags from sample to be tested

RNA fragment hybridizes with DNA on GeneChip
Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow.
diauxic shift timecourse: 0 hr  diauxic shift timecourse: 20.5 hr

Microarray Databases

- Stanford Microarray Database
  - Experiments: 66571
  - Public Experiments: 12596
  - Spots: 1983883115
  - Users: 1633
  - Labs: 319
  - Organisms: 50
  - Publications: 350

<table>
<thead>
<tr>
<th>Amino Acids</th>
<th>(nonpolar = hydrophobic)</th>
<th>Polarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>ala A</td>
<td>N</td>
</tr>
<tr>
<td>Arginine</td>
<td>arg R</td>
<td>+</td>
</tr>
<tr>
<td>Asparagine</td>
<td>asn N</td>
<td>P</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>asp D</td>
<td>-</td>
</tr>
<tr>
<td>Cysteine</td>
<td>cys C</td>
<td>P</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>glu E</td>
<td>-</td>
</tr>
<tr>
<td>Glutamine</td>
<td>gln Q</td>
<td>P</td>
</tr>
<tr>
<td>Glycine</td>
<td>gly G</td>
<td>N</td>
</tr>
<tr>
<td>Histidine</td>
<td>his H</td>
<td>+</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>ile I</td>
<td>N</td>
</tr>
<tr>
<td>Leucine</td>
<td>leu L</td>
<td>N</td>
</tr>
<tr>
<td>Lysine</td>
<td>lys K</td>
<td>+</td>
</tr>
<tr>
<td>Methionine</td>
<td>met M</td>
<td>N</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>phe F</td>
<td>N</td>
</tr>
<tr>
<td>Proline</td>
<td>pro P</td>
<td>N</td>
</tr>
<tr>
<td>Serine</td>
<td>ser S</td>
<td>P</td>
</tr>
<tr>
<td>Threonine</td>
<td>thr T</td>
<td>P</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>trp W</td>
<td>N</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>tyr Y</td>
<td>P</td>
</tr>
<tr>
<td>Valine</td>
<td>val V</td>
<td>N</td>
</tr>
</tbody>
</table>

GGUGCGCGUUAU

GARY
Amino acids
Consists of a central carbon atom $C_\alpha$ which is bonded to an amino group and a carboxyl group and a side-chain

The backbone in proteins
The backbone is the sequence of amino groups, $C_\alpha$, and carboxyl groups

Proteins differ only in the number of amino acids linked together, and the sequential order in which these amino acids occur
Protein

- 20 amino acid types
- Folding

Primary protein structure is a sequence of a chain of amino acids.

Secondary protein structure occurs when the sequence of amino acids are linked by hydrogen bonds.

Tertiary protein structure occurs when certain attractions are present between alpha helices and pleated sheets.

Quaternary protein structure is a protein consisting of more than one amino acid chain.

Image adapted from: National Human Genome Research Institute.
Protein Analysis

- Broad area of bioinformatics includes:
  - Sequence
  - Structure
  - Function

- Focus today on:
  - Finding Motifs
  - Classification
  - Prediction
Primary Data Sources

- Sequence – pdb, swissProt
- Structure – cath, dssp
- Function - cath, scop
- Experimentally derived from a lab or group of labs (e.g. NMR data for membrane spanning proteins)
Quantitative Structure-Activity Relationships

- Molecule/activity set
- 10s-100s of features = (hydrophobicity, electronic effects, steric effects, etc.)
- Training / testing / validation sets
- Reduce the number of features
- Predictive model for future compound discovery
- Better understanding of true biological mechanism of action
Phylogenetics

- Determining the history of life on Earth using sequence information or other characteristics/features
- Develop a tree-like representation
- Factorial increase in the number of possible trees as the number of sequences/features increases
- Better search algorithms for the space of tree representations for resolution of a “true” tree
Data Pitfalls

- Bias due to homology
- Dirty data due to unknowns
- Not enough examples to train from
- Outliers
- Out of date
- Garbage in, garbage out
Predictors and Classifiers

- **Predictors** – given an unknown data sample provide a measure of confidence that it belongs to the set the model was constructed to represent.

- **Classifiers** – given a set of heterogeneous data separate the data into classes.
  - Supervised – trained on a set of known examples
  - Unsupervised – number of classes is not known in advance (also referred to as clustering)
Clustering

- Clustering is the classification of similar objects into different groups.

- More precisely - the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure.
Measuring Performance

- TP(t), FP(t), TN(t), FN(t), where $t = \text{threshold}$

- **Specificity**
  - $\frac{TN}{TN+FP}$

- **Sensitivity**
  - $\frac{TP}{TP+FN}$

- **False Positive Rate**
  - $1 - \text{specificity} = \frac{FP}{FP+TN}$
Measuring Performance

- **Correlation coefficient**
  \[
  \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}}
  \]

- **Receiver operating characteristic**
  - Sensitivity/specificity plot for an experiment
  - Measure the area under the curve

![Bivariate Scattergram](image-url)
When to Employ an Evolutionary Algorithm

- Large search space with many local optima
- Neither exact algorithms nor approximation algorithms feasible
- Applications where current solutions rely on heuristics
- Dynamic processes

Examples in bioinformatics:
- multiple sequence alignment
- structure prediction
- clustering expression data
- phylogeny (using parsimony)
- parameter estimation in hidden Markov models
- finding gene networks
Problem Dependent Application

- Each problem requires its own
  - Representation
    - Particularly important in bioinformatics
  - Variation operators
  - Rates of mutation/recombination
  - Performance index
Gene Expression

- Class prediction through evolved classifiers
  - which genes are most correlated with known cell types/disease phenotype

- Class discovery through evolutionary computation
  - how many cell types are truly represented in the gene expression data?
Optimizing Neural Networks Using Evolutionary Computation

- Weights and biases
- Connections
- Topology
- Processing elements
Strategy for Broad-Spectrum Drug Discovery

- Search nucleotide sequences for conserved RNA structures/drug targets with broad spectrum anti-bacterial or anti-viral activity
Exhaustive search for common RNA structures infeasible

>10^5 hits
Organism A

>10^5 hits
Organism B

>10^5 hits
Organism C

Structure A_{134}
Structure B_{56}
Structure C_{278}

Bin #1

Structure A_{78}
Structure B_{9}
Structure C_{3567}

Bin #2

...

Structure A_{x}
Structure B_{y}
Structure C_{z}

Bin #N

→ Scoring

← Variation

↓ Selection
Signal Recognition Particle – Domain IV

- SRP targets signal peptide-containing proteins to plasma membranes (prokaryotes) or endoplasmic reticulum (eukaryotes)

- Domain IV is essential, known binding site for protein component of SRP

- Highly conserved, found over a wide phylogenetic distance
<table>
<thead>
<tr>
<th>Experiment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organism</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>20</td>
<td>136</td>
<td>418</td>
<td>903</td>
<td>903</td>
<td>903</td>
</tr>
<tr>
<td><em>A. fulgidus</em></td>
<td>15</td>
<td>69</td>
<td>200</td>
<td>724</td>
<td>724</td>
<td>724</td>
</tr>
<tr>
<td><em>B. subtilis</em></td>
<td>12</td>
<td>62</td>
<td>374</td>
<td>520</td>
<td>520</td>
<td>520</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>14</td>
<td>45</td>
<td>258</td>
<td>523</td>
<td>523</td>
<td>523</td>
</tr>
<tr>
<td><em>M. voltae</em></td>
<td>11</td>
<td>30</td>
<td>121</td>
<td>315</td>
<td>315</td>
<td>315</td>
</tr>
<tr>
<td><em>S. pyogenes</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td>57295</td>
</tr>
<tr>
<td><em>S. aureus</em></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>13591</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>72</td>
<td>342</td>
<td>1371</td>
<td>2985</td>
<td>60280</td>
<td>16576</td>
</tr>
</tbody>
</table>
Signal Recognition Particle – Domain IV

<table>
<thead>
<tr>
<th>Exp.</th>
<th>Possible Bins</th>
<th>P</th>
<th>O</th>
<th>G</th>
<th>Time (min.)</th>
<th>Fraction Evaluated</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$5.5 \times 10^5$</td>
<td>80</td>
<td>40</td>
<td>3</td>
<td>2</td>
<td>$1.7 \times 10^{-2}$</td>
</tr>
<tr>
<td>2</td>
<td>$7.9 \times 10^8$</td>
<td>80</td>
<td>40</td>
<td>7</td>
<td>4</td>
<td>$2.8 \times 10^{-5}$</td>
</tr>
<tr>
<td>3</td>
<td>$9.8 \times 10^{11}$</td>
<td>80</td>
<td>40</td>
<td>27</td>
<td>13</td>
<td>$8.8 \times 10^{-8}$</td>
</tr>
<tr>
<td>4</td>
<td>$5.6 \times 10^{13}$</td>
<td>80</td>
<td>40</td>
<td>27</td>
<td>12</td>
<td>$5.9 \times 10^{-11}$</td>
</tr>
<tr>
<td>5</td>
<td>$3.2 \times 10^{18}$</td>
<td>200</td>
<td>100</td>
<td>25</td>
<td>90</td>
<td>$1.5 \times 10^{-13}$</td>
</tr>
<tr>
<td>6</td>
<td>$7.6 \times 10^{17}$</td>
<td>200</td>
<td>100</td>
<td>13</td>
<td>41</td>
<td>$3.4 \times 10^{-13}$</td>
</tr>
</tbody>
</table>

Experiments 1-4

<table>
<thead>
<tr>
<th>gi</th>
<th>38795</th>
<th>192</th>
<th>24</th>
<th>gcc cagg  ccc ggaa ggg agca  gcc</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>216348</td>
<td>153</td>
<td>24</td>
<td>tgt cagg  tcc ggaa gga agca  gca</td>
</tr>
<tr>
<td>gi</td>
<td>42758</td>
<td>204</td>
<td>24</td>
<td>ggt cagg  tcc ggaa gga agca  gcc</td>
</tr>
<tr>
<td>gi</td>
<td>177793</td>
<td>308</td>
<td>24</td>
<td>gcc cagg  tcg gaaa cgg agca  ggt</td>
</tr>
<tr>
<td>gi</td>
<td>150042</td>
<td>310</td>
<td>26</td>
<td>ccg ccagg ccc ggaa ggg agcaag cgg</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>G-C</td>
<td></td>
</tr>
<tr>
<td>C-G</td>
<td></td>
</tr>
<tr>
<td>U-G</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td></td>
</tr>
<tr>
<td>C-G</td>
<td></td>
</tr>
<tr>
<td>C-G</td>
<td></td>
</tr>
<tr>
<td>G-U</td>
<td></td>
</tr>
</tbody>
</table>
Signal Recognition Particle – Domain IV

Experiment 5

- gi|38795| 192 24 gcc cagg ccc ggaa ggg agca ggc
- gi|216348| 153 24 tgt cagg tcc ggaa gga agca gca
- gi|42758| 204 24 ggt cagg tcc ggaa gga agca gcc
- gi|177793| 308 24 gcc cagg tcg gaaa cgg agca ggt
- gi|150042| 310 26 ccg ccagg ccc ggaa ggg agcaa cgg
- gb|AE004092| 190360 24 ggt cagg gga ggaa tcc agca gcc

Experiment 6

- gi|38795| 192 24 gcc cagg ccc ggaa ggg agca ggc
- gi|216348| 153 24 tgt cagg tcc ggaa gga agca gca
- gi|42758| 204 24 ggt cagg tcc ggaa gga agca gcc
- gi|177793| 308 24 gcc cagg tcg gaaa cgg agca ggt
- gi|150042| 310 26 ccg ccagg ccc ggaa ggg agcaa cgg
- gi|15922990| 525890 24 tgt cagg tcc tgac gga agca gca
Transcription Factor Binding Site (TFBS) Discovery

- Use evolutionary computation to search for TFBSs of co-expressed genes

- Identify known TFBS motifs as well as putative, previously unknown motifs that serve as promoters or enhancers

- Follow up with experimental validation
Discovery of TFBSs using EC

References – Bioinformatics Books

References

- Fogel GB “Microarray Data Mining with Evolutionary Computation,” in *Evolutionary Computation in Data Mining*, (A. Ghosh and LC Jain eds.) Springer, 2005.
Tutorial on Evolutionary Computation in Bioinformatics: Part II

Kay C. Wiese
School of Computing Science
Simon Fraser University
Canada
Chair: IEEE CIS Bioinformatics and Bioengineering Technical Committee

Gary B. Fogel
Vice President
Natural Selection, Inc.
9330 Scranton Rd., Suite 150
San Diego, CA 92121 USA
Overview

- Computational Intelligence (CI) in Bioinformatics
- RNA Structure and Prediction
- Designing Algorithms Inspired by Nature: Evolutionary Computation – Successes in the RNA Domain with RnaPredict
- A comparison with known structures and mfold
- jViz.Rna – A Dynamic RNA Visualization Tool
- Conclusions
Computational and Design Methods Used in Bioinformatics

- Algorithm Design including:
  - Dynamic Programming
  - Heuristic Search
  - Computational Intelligence Methods: Evolutionary Computation, Simulated Annealing, Neural Networks, Fuzzy Systems

- Graphics and Visualization
  - Designing models for sequence or structure information
  - 2D and 3D visualization of structures
  - Systems design for input, output and manipulation
  - Design of interactive input and output technology
Successful Applications of CI in Bioinformatics

- Sequence Alignment
  - SAGA: An Evolutionary Algorithm (EA) for Sequence Alignment (Notredame et al.)

- RNA Structure Prediction
  - Evolutionary Algorithms (RnaPredict and PRnaPredict (Wiese et al., Shapiro et al., vanBatenburg et al.))
  - Simulated Annealing
Successful Applications of CI in Bioinformatics, cont.

- Protein Structure Prediction (Searching the Protein Conformational Space)
  - Evolutionary Algorithms
  - Simulated Annealing
  - Knowledge Based Methods

- Protein-Protein Interaction
  - Do two proteins interact? Where? How?
  - Searching Conformational Space for Docking (EAs)
Successful Applications of CI in Bioinformatics, cont.

- Identify Coding Regions in DNA
  - Evolved Artificial Neural Network for Gene Identification (Fogel et al.)

- Excellent Overview of EA approaches in Bioinformatics
  - *Computational Intelligence in Bioinformatics*, Fogel et al., IEEE Press (due Dec, 2007)
RNA Folding – A case study

- RNA is involved in transcription and translation: making proteins
- Other roles include regulatory, catalytic and structural roles, also in combination with proteins
- RNA sequences are determined using high throughput sequencing machines
RNA Folding: Why should we care?

- Why study/predict RNA structure?
- The structure of RNA molecules largely determines their function in the cell.
- Preservation of structure can be used to understand evolutionary processes.
- Knowing the structure or shape can be used to understand genetic diseases and to design new drugs.
- RNA secondary structure is formed by a natural folding process in which chemical bonds between certain so called canonical base pairs are formed.
RNA Folding

- The canonical base pairs are GC, AU, GU, and mirrors CG, UA, UG

  e.g.

  $\begin{array}{c}
  5' \\
  A \\
  U \\
  3'
  \end{array}$

- Finding all canonical base pairs is simple, but which ones will actually form bonds?
RNA Folding

- While the sequence “folds” back onto itself it forms the secondary structure
RNA Secondary Structure Elements

Note: the same sequence may produce many different, overlapping helices
RNA Double Helix Model

- A helix consists of at least three consecutive canonical base pairs.
- The helix can only form if the sequence or loop connecting the two strands is at least 3 nucleotides long.
Which helices are possible?
Which helices are possible?
RNA Folding by Energy Minimization

- RNA molecules are stabilized by the formation of these helices (through base pair bonds).
- How do we know which helices will form?
- RNA molecules will fold into a minimum energy state. This minimum can be a local one.
- The free energy of a structure is determined by evaluating the thermodynamic model that is associated with the current structure.
RNA Thermodynamics

- Energies for various RNA substructures can be determined experimentally and are associated with thermodynamic parameters.
- Thermodynamic models can consider bonding energies, stacking energies, and looping energies.
- Our work employs two stacking energy models
  - Individual Nearest Neighbor (INN) (Borer, Freier, Sugimoto, He)
  - Individual Nearest Neighbor Hydrogen Bond (INN-HB) (Xia et al. 1998)
Free Energy Minimization
Stacking Energies

- Free energy ($\Delta G$) is reduced as base pairs are formed
- Two helices with same base pairs can have different $\Delta G$
- $\Delta G$ contribution of a base pair depends on
  - Position in helix
  - Proximity to other base pairs
- Both INN and INN-HB model nearest-neighbors, terminal mismatches, dangling ends, helix initiation, and helix symmetry
Stacking Energies (INN and INN-HB)

Suppose duplex:

\[
\Delta G(\text{duplex}) = \Delta G_{\text{initiation}} + \Delta G \left( \begin{array}{c} \text{G} \\ \text{A} \end{array} \right) + \Delta G \left( \begin{array}{c} \text{A} \\ \text{U} \end{array} \right) + \Delta G \left( \begin{array}{c} \text{U} \\ \text{A} \end{array} \right) + \Delta G_{\text{symmetry}}
\]
Energy Minimization

The free energy $\Delta G(S)$ of the entire structure is given by:

$$\Delta G(S) = \sum_{h \in S} \Delta G(h)$$

where $\Delta G(h)$ is the free energy of an individual helix according to the thermodynamic model in use.
Approach - Find All Pairs

1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by "stacking" pairs
3. Add helix to set H of all potential helices
Build All Helices

1. Find all canonical base pairs
2. Attempt to grow each pair \((i,j)\) into a helix by “stacking” pairs
3. Add helix to set \(H\) of all potential helices
Build All Helices

1. Find all canonical base pairs
2. Attempt to grow each pair(i, j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices.
Build All Helices

1. Find all canonical base pairs
2. Attempt to grow each pair \((i,j)\) into a helix by “stacking” pairs
3. Add helix to set \(H\) of all potential helices
Build All Helices

1. Find all canonical base pairs
2. Attempt to grow each pair \((i,j)\) into a helix by “stacking” pairs
3. Add helix to set \(H\) of all potential helices.
Build All Helices

1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by "stacking" pairs
3. Add helix to set H of all potential helices
Add New Helix

1. Find all canonical base pairs
2. Attempt to grow each pair(i,j) into a helix by “stacking” pairs
3. Add helix to set H of all potential helices
RNA Helices

- Must have at least 3 “stacked” base pairs
- Sequence or loop connecting the two strands must be at least 3 nucleotides long
- Store this new helix in a set H
Assembling a Structure $S$

- From the set $H$ of all helices, select a subset $S$ such that:
  - Sum of free energies of all $h$ in $S$ is minimized
  - No helices $h$ in $S$ share bases

$$H = \begin{cases}
\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\}
\end{cases}$$

$$S = \begin{cases}
\{1, 4, 7, 8, 11\}
\end{cases}$$
The energy function $E$ is determined by the current thermodynamic model and minimized

$$\sum_{h \in S} \Delta G(h) = \min$$

Challenge: There are $2^{|H|}$ sub-sets of $H$

Solution: Probabilistic Approaches, such as Evolutionary Algorithms, Monte Carlo, Simulated Annealing
Objectives

- To design a novel Evolutionary Algorithm to predict secondary RNA structure (RnaPredict)
- To evaluate the algorithm including convergence behavior and population dynamics
- To suggest several improvements to the algorithm
- To study the effect of different selection and reproduction techniques on the EA
- To compare the outcome (predicted structures) with known structures and other approaches such as Nussinov and mfold
Initialize a population of chromosomes; Evaluate the chromosomes in the population;

while (stopping criteria not reached) do
    for i=1 to size_of(population)/2 do
        select 2 parent chromosomes;
        apply crossover operator to them;
        apply mutation operator to them;
        evaluate the new chromosomes;
        insert them into $g_{next}$;
        i = i + 1;
    endfor
    update stopping criteria;
endwhile
Representing Structure S in the Algorithm

- A permutation P of set H is used to represent the structure in the EA

\[ P_x = \{8, 13, 3, 9, 12, 6, 7, 2, 5, 4, 11, 10, 1\} \]

\[ P_y = \{13, 3, 10, 4, 5, 11, 9, 8, 1, 12, 6, 7, 2\} \]
Decoding the Permutation

- When an individual is decoded, each helix in the permutation is iterated through

\[ S_x = \{8, 12, 6, 4, 1\} \]

\[ S_y = \{13, 3, 10, 4, 8, 1, 6, 7\} \]

- Only helices which do not conflict are scored and placed in the final structure
Experiments and Results

- Tested several sequences with both known and unknown structures
- 4 selection/replacement strategies
- 2 representations and 9 X-over operators
- over 100 combinations of $P_c$ and $P_m$
- Population size = 700
- 30 independent runs with different random seeds
- 785 nt human mRNA ...
Results – Roulette Wheel Selection

785 nt human mRNA, 10480 Possible Helices, \( P_m = 5\% \), \( P_c = 70\% \), pop_size = 700, CX, STDS, no elitism
Results: Keep-Best Reproduction

- Faster convergence
- Better solution quality
- Works well with smaller population sizes
- Very robust over a wide range of operator probabilities
Results – KBR, cont.

785 nt human mRNA, 10480 Possible Helices, $P_m=80\%$, $P_c=70\%$

$\text{pop\_size} = 700$, CX, KBR, 1-elitism
What about the structures?

- So far we have focused on understanding the factors that control:
  - the convergence speed of the algorithm
  - the effectiveness of the algorithm to find low energy structures

- What about the actual structures?
  - Need to know how close our predicted structures match real structures in nature
## Predicted Structures vs. Real Structures

<table>
<thead>
<tr>
<th></th>
<th>Bonding Model</th>
<th>Stacking Model (INN-HB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly predicted base pairs</td>
<td>&lt;30%</td>
<td>71.1%</td>
</tr>
<tr>
<td>Best</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Best canonical base pairs only</td>
<td>41.3%</td>
<td>93.1%</td>
</tr>
</tbody>
</table>

_Haloarcula marismortui – 122nt_
What about the quality of the structures?

- How much does the previous quantitative measure (base pair overlap) tell us?
- What do the structures look like?
Visualization of RNA secondary structure – *jViz.Rna*

We have developed a tool (*jViz.Rna*) to visualize RNA that could:

- Be separate from the prediction tool
- Handle and display pseudo-knots
- Have dynamic output for further manipulation by the user
- Allow for easy comparison of two structures
- Allow for quantitative comparison of two structures
- Allow for saving the output as high quality graphics in a standard format (.gif)
jViz.Rna – Feynman diagram
Known Structure

*Saccharomyces cerevisiae* (Baker’s Yeast) 118 nt

**Known**
**Predicted**
**Overlap**
jViz.Rna – Predicted Structure

Saccharomyces cerevisiae (Baker's Yeast) 118 nt

Known
Predicted
Overlap
**jViz.Rna** – Comparison of Predicted vs. Known Structure

*Saccharomyces cerevisiae* (Baker's Yeast) 118 nt

Known
Predicted
Overlap
jViz.Rna – Dot Plot

Saccharomyces cerevisiae (Baker's Yeast) 118 nt
$j\text{Viz.Rna} - \text{Classical Structure (known)}$

$\text{Saccharomyces cerevisiae}$
(Baker's Yeast) 118 nt
**jViz.Rna** – Classical Structure (predicted)

*Saccharomyces cerevisiae*
(Baker's Yeast) 118 nt
jViz.Rna – Overlaying predicted and known structure

Saccharomyces cerevisiae
(Baker's Yeast) 118 nt
RNA Base Pairing

- Canonical
  - G-C
  - A-U
  - G-U

- Non-canonical
  - G-A
  - A-A
  - C-U
  - U-U
  - and others…
Quantitative Overlap (known vs. predicted structure)

- Known structure of Baker’s Yeast has a total of 37 bps
- Our method *RnaPredict* predicts 33 of those (89.2%)
- Known structure contains 2 C-U pairs which cannot be predicted with the current model: 33/35 were found (94.3%)
- Without the C-U pairs the existing helix of size 3 would only be size 2 and could thus not be predicted
- Of the bps and helices that our model can predict, it found 100%, hence the search engine has a very high accuracy
A comparison with *Nussinov DPA*

- Nussinov is a simplistic DPA for RNA secondary structure prediction that works on the principle of base pair maximization
- Modifications made to emulate bp weights at $G-C = 3$, $A-U = 2$, and $G-U = 1$
- Used as a base line for our comparisons
S. cerevisiae via Nussinov

Predicted

Overlap
**Best Nussinov vs. best Correct BP EA run**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>37</td>
<td>17</td>
<td>6</td>
<td>28</td>
<td>33</td>
<td>75.7%</td>
<td>89.2%</td>
</tr>
<tr>
<td><em>H. marismortui</em></td>
<td>38</td>
<td>37</td>
<td>17</td>
<td>8</td>
<td>16</td>
<td>21.1%</td>
<td>71.1%</td>
</tr>
<tr>
<td><em>H. rubra</em></td>
<td>138</td>
<td>174</td>
<td>88</td>
<td>31</td>
<td>71</td>
<td>22.5%</td>
<td>51.4%</td>
</tr>
<tr>
<td><em>D. virilis</em></td>
<td>233</td>
<td>291</td>
<td>168</td>
<td>29</td>
<td>66</td>
<td>12.4%</td>
<td>28.3%</td>
</tr>
<tr>
<td><em>X. laevis</em></td>
<td>251</td>
<td>286</td>
<td>158</td>
<td>47</td>
<td>100</td>
<td>18.7%</td>
<td>39.8%</td>
</tr>
</tbody>
</table>
## Best Nussinov vs. best Correct BP EA run

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. lagunensis</em></td>
<td>113</td>
<td>142</td>
<td>59</td>
<td>30</td>
<td>73</td>
<td>26.5%</td>
<td>64.6%</td>
</tr>
<tr>
<td><em>A. griffini</em></td>
<td>131</td>
<td>166</td>
<td>78</td>
<td>48</td>
<td>79</td>
<td>36.6%</td>
<td>60.3%</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>189</td>
<td>281</td>
<td>146</td>
<td>26</td>
<td>56</td>
<td>13.8%</td>
<td>29.6%</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>266</td>
<td>309</td>
<td>135</td>
<td>33</td>
<td>92</td>
<td>12.4%</td>
<td>34.6%</td>
</tr>
<tr>
<td><em>S. acidocaldarius</em></td>
<td>468</td>
<td>395</td>
<td>288</td>
<td>187</td>
<td>159</td>
<td><strong>40.0%</strong></td>
<td>34.0%</td>
</tr>
</tbody>
</table>
A comparison with *mfold*

- *mfold* is the most cited and widely used RNA secondary structure prediction algorithm (based on DP)
- Developed by Zuker et al.
- Basic algorithm first introduced in 1981
- Since then refined continually until today (newest version published in 2003)
- Considered the gold standard of RNA secondary structure prediction
Overall Best Correct BP *mfold* vs. EA result

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em></td>
<td>37</td>
<td>8</td>
<td>6</td>
<td>33</td>
<td>33</td>
<td>89.2%</td>
<td>89.2%</td>
</tr>
<tr>
<td><em>H. marismortui</em></td>
<td>38</td>
<td>5</td>
<td>17</td>
<td>29</td>
<td>27</td>
<td>76.3%</td>
<td>71.1%</td>
</tr>
<tr>
<td><em>H. rubra</em></td>
<td>138</td>
<td>127</td>
<td>88</td>
<td>49</td>
<td>71</td>
<td>35.5%</td>
<td>51.4%</td>
</tr>
<tr>
<td><em>D. virilis</em></td>
<td>233</td>
<td>199</td>
<td>168</td>
<td>37</td>
<td>66</td>
<td>15.9%</td>
<td>28.3%</td>
</tr>
<tr>
<td><em>X. laevis</em></td>
<td>251</td>
<td>157</td>
<td>158</td>
<td>92</td>
<td>100</td>
<td>36.7%</td>
<td>39.8%</td>
</tr>
</tbody>
</table>
Overall Best Correct BP *mfold* vs. EA result

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. lagunensis</em></td>
<td>113</td>
<td>68</td>
<td>59</td>
<td>60</td>
<td>73</td>
<td>53.1%</td>
<td>64.6%</td>
</tr>
<tr>
<td><em>A. griffini</em></td>
<td>131</td>
<td>105</td>
<td>78</td>
<td>67</td>
<td>79</td>
<td>51.1%</td>
<td>60.3%</td>
</tr>
<tr>
<td><em>C. elegans</em></td>
<td>189</td>
<td>177</td>
<td>146</td>
<td>40</td>
<td>56</td>
<td>21.2%</td>
<td>29.6%</td>
</tr>
<tr>
<td><em>H. sapiens</em></td>
<td>266</td>
<td>163</td>
<td>135</td>
<td>95</td>
<td>92</td>
<td>35.7%</td>
<td>34.6%</td>
</tr>
<tr>
<td><em>S. acidocaldarius</em></td>
<td>468</td>
<td>233</td>
<td>288</td>
<td>261</td>
<td>159</td>
<td>55.8%</td>
<td>34.0%</td>
</tr>
</tbody>
</table>
2 extra base pairs predicted by *mfold*

Saccharomyces cerevisiae (Baker's Yeast) 118 nt
Pseudo-knots

- 2 base pairs (i, j) and (i’, j’) are pseudoknotted if i < i’ < j < j’… a 3D interaction.
- Occurrence of pseudo-knots is rather rare, but structurally significant
- Typically, longer sequences have a higher probability of having pseudo-knots
- How can we display them?
  - Arc and circular diagrams can display pseudo-knots but do not work well for longer sequences
  - Classical structure? Existing tools such as RnaViz do a poor job, example...
H. rubra

RnaViz
jViz.Rna – Classical Structure and Pseudo-knots

Known

H. rubra
Conclusions – RNA Folding

- Excellent convergence behavior of EA
- Best results achieved with Keep-Best Reproduction (both efficiency and quality of structures increase substantially)
- Very high accuracy of prediction for shorter sequences
- EA is able to work with the “fuzziness” of the noisy thermodynamic model
Conclusion – RNA Folding cont.

- Outperforms Nussinov DP in all but one case by a wide margin
- Outperforms mfold on several sequences, despite mfold’s use of a very sophisticated thermodynamic model
- mfold also cannot predict non-canonical base pairs
Conclusion – jViz.Rna

- jViz.Rna is a platform independent visualization tool for RNA structure
- jViz.Rna applications include:
  - study RNA structure
  - evaluation of RNA folding algorithms
Venues of Interest

- IEEE Symposium on Computational Intelligence in Bioinformatics and Comp. Biology (CIBCB) – www.cibcb.org
- Special Session on EC in Bioinformatics and Comp. Biology at IEEE Congress on Evolutionary Computation.
- IEEE/ACM Transactions on Comp. Biology and Bioinformatics
- IEEE Transactions on NanoBioscience
- IEEE Transactions on Evolutionary Computation
Acknowledgements

- Simon Fraser University
- Natural Science and Engineering Research Council of Canada
- Canada Foundation for Innovation
- Andrew Hendriks – Presentation, Research and Programming
- Edward Glen – Visualization of structures
- Alain Deschenes – Research and Programming
- Natural Selection, Inc.
- National Science Foundation
- National Institutes of Health
- Dana Weekes, Mars Cheung – Research and Programming
- Institute of Electrical and Electronics Engineers IEEE and IEEE/CIS
- L. Gwenn Volkert, Dan Ashlock, Clare Bates Congdon
Questions?
References


- Pedersen, Jan T. and Moult, John (1996), "Genetic algorithms for protein structure prediction", *Current Opinion in Structural Biology*, 6 (2), 227-231


Michael Zuker: [http://bioinfo.math.rpi.edu](http://bioinfo.math.rpi.edu)