Métodos de Inteligencia Artificial en Biología Computacional
Artificial Intelligence Methods in Computational Biology

José Santos
Universidad de A Coruña
jose.santos@udc.es
Computational Biology. Definition

**Bioinformatics**: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

**Computational Biology**: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology.

NIH Working Definitions of Bioinformatics and Computational biology, 2000:

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the previous definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.
Computational Biology. Definition

Other definition:

**Computational biology**: the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. *It is about science.*

**Bioinformatics**: the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. *It is about engineering.*

(Russ Altman, Stanford University).
Computational Biology. Areas of research
Protein structure prediction and protein folding modeling with cellular automata
Proteins

“A b initio” prediction

(a) Primary structure

(b) Secondary structure

(c) Tertiary structure

(d) Quaternary structure

Hydrogen bonds between amino acids at different locations in polypeptide chain

Pleated sheet

β polypeptide

Heme group
Proteins

“Ab initio” prediction
## Proteins - PDB (Protein Data Bank)

<table>
<thead>
<tr>
<th>Experimental Method</th>
<th>Proteins</th>
<th>Nucleic Acids</th>
<th>Protein/Nucleic Acid complexes</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray diffraction</td>
<td>95636</td>
<td>1694</td>
<td>4817</td>
<td>4</td>
<td>102151</td>
</tr>
<tr>
<td>NMR</td>
<td>9840</td>
<td>1135</td>
<td>231</td>
<td>8</td>
<td>11214</td>
</tr>
<tr>
<td>Electron microscopy</td>
<td>666</td>
<td>29</td>
<td>227</td>
<td>0</td>
<td>922</td>
</tr>
<tr>
<td>Hybrid</td>
<td>83</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>89</td>
</tr>
<tr>
<td>Other</td>
<td>170</td>
<td>4</td>
<td>6</td>
<td>13</td>
<td>193</td>
</tr>
<tr>
<td><strong>Total:</strong> 106293</td>
<td>2865</td>
<td>5283</td>
<td>26</td>
<td></td>
<td><strong>114569</strong></td>
</tr>
</tbody>
</table>
Proteins.

HP Model

**NONPOLAR, HYDROPHOBIC**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>R Groups</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>H - CH₂ - COO⁻</td>
<td>89</td>
</tr>
<tr>
<td>Valine</td>
<td>HO - CH₂ - COO⁻</td>
<td>117</td>
</tr>
<tr>
<td>Leucine</td>
<td>OH - CH₃ - COO⁻</td>
<td>131</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>HS - CH₂ - COO⁻</td>
<td>131</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>HO - C₆H₅ - COO⁻</td>
<td>181</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>NH₂ - C - CH₂ - COO⁻</td>
<td>204</td>
</tr>
<tr>
<td>Methionine</td>
<td>NH₂ - C - CH₂ - CH₃ - S - CH₃</td>
<td>149</td>
</tr>
<tr>
<td>Proline</td>
<td>+ NH₃ - CH₂ - (CH₃)₂ - CH⁺ - COO⁻</td>
<td>115</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>NH₃ - CH₂ - C - COO⁻</td>
<td>133</td>
</tr>
<tr>
<td>Glutamine acid</td>
<td>N₃⁺ - CH₂ - CH₂ - C - COO⁻</td>
<td>147</td>
</tr>
</tbody>
</table>

**POLAR, UNCHARGED**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>R Groups</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine</td>
<td>H - CH₂ - NH₃⁺</td>
<td>75</td>
</tr>
<tr>
<td>Serine</td>
<td>HO - CH₂ - NH₃⁺</td>
<td>105</td>
</tr>
<tr>
<td>Threonine</td>
<td>OH - CH₃ - NH₃⁺</td>
<td>119</td>
</tr>
<tr>
<td>Cysteine</td>
<td>HS - CH₂ - NH₃⁺</td>
<td>121</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>HO - C₆H₅ - NH₃⁺</td>
<td>181</td>
</tr>
<tr>
<td>Asparagine</td>
<td>NH₂ - C - CH₂ - COO⁻</td>
<td>132</td>
</tr>
<tr>
<td>Glutamine</td>
<td>NH₂ - C - NH₂ - (CH₃)₂ - CH₃ - NH₃⁺</td>
<td>146</td>
</tr>
<tr>
<td>Lysine</td>
<td>NH₃⁺ - CH₂ - CH₂ - C - COO⁻</td>
<td>146</td>
</tr>
<tr>
<td>Arginine</td>
<td>NH₂ - C - NH₂ - (CH₃)₂ - CH₃ - NH₃⁺</td>
<td>174</td>
</tr>
<tr>
<td>Histidine</td>
<td>HN⁺ - CH₂ - CH₂ - COO⁻</td>
<td>155</td>
</tr>
</tbody>
</table>
**HP model**

- Amino acids are classified in:
  - **H** (hydrophobic): low propensity to be in contact with water, tendency to be buried inside the protein core
  - **P** (polar): tendency to be in the protein surface in contact with water

- Each protein is represented as a chain:
  
  HPHPPHHPHPPHPHPPHPHPPHPH

---

Grid 2D

Grid 3D

Benchmarks

- **Opt - 23**
- **Opt - 21**
- **Opt - 42**
The energy reflects the fact that the hydrophobic amino acids have a propensity to form a hydrophobic inner core.
Protein Conformation Representation: relative moves

Relative moves: F, L, R (Grid 2D)
**Algorithm 1** Differential Evolution Algorithm.

1: Initialize the population randomly
2: repeat
3: for all individual $x$ in the population do
4: Let $x_1, x_2, x_3 \in$ population, randomly obtained \{$x_1, x_2, x_3, x$ different from each other.\}
5: Let $R \in \{1, \ldots, n\}$, randomly obtained \{$n$ is the length of the chain.\}
6: for $i = 1$ to $n$ do
7: Pick $r_i \in U(0,1)$ uniformly from the open range (0,1).
8: if $(i = R) \vee (r_i < CR)$ then
9: \hspace{1cm} $y_i \leftarrow x_{1i} + F(x_{2i} - x_{3i})$
10: else \hspace{1cm} $y_i = x_i$
11: end if
12: end for
13: end for\{y = [y_1, y_2 \ldots y_n]$ is a new generated candidate individual\}
14: if $f(y) < f(x)$ then
15: Replace individual $x$ by $y$
16: end if
17: end for
18: until termination criterion is met
19: return $z \in$ population \(\forall t \in$ population, $f(z) \leq f(t)$

In our application: $F:0.9$

**CR:** 0.9

**X1** selected with tournament
Differential Evolution. Encoding

We used relative coordinates. Three movements in 2D: (F)orward, (R)ight and (L)eft

\[
\begin{array}{cccccc}
0.24 & -0.33 & 2.44 & -1.25 & 0.18 & \ldots
\end{array}
\]

\[
\begin{array}{cccccc}
F & F & R & L & F
\end{array}
\]

movement  L if \( X_{ij} \) [\( \alpha, \beta \)]

F if \( X_{ij} \) [\( \beta, \delta \)]

R if \( X_{ij} \) [\( \delta, \gamma \)]

\( \alpha < \beta < \delta < \gamma \) arbitrary constants in \( \mathbb{R} \) (\( \alpha=-3, \beta=-1, \delta=1, \gamma=3 \))
Repair process

**Absolute moves procedure**: tries to maintain the relative conformation of the rest of the chain

**Cartesian coordinates procedure**: tries to obtain legal conformations searching for a similar one in the Cartesian space
### Some results

Comparison of results with the benchmark sequences

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>-9</td>
<td>-9 (30492)</td>
<td>-9,-9 (3584, 6362)</td>
<td>-9,-9</td>
<td>-9</td>
</tr>
<tr>
<td>S2</td>
<td>-9</td>
<td>-9 (30491)</td>
<td>-9,-9 (5806, 9292)</td>
<td>-9,-9</td>
<td>-9</td>
</tr>
<tr>
<td>S3</td>
<td>-8</td>
<td>-8 (20400)</td>
<td>-8,-8 (7061, 18828)</td>
<td>-8,-8</td>
<td>-8</td>
</tr>
<tr>
<td>S4</td>
<td>-14</td>
<td>-14 (301339)</td>
<td>-14,-14 (45793, 92579)</td>
<td>-14,-13.96</td>
<td>-14</td>
</tr>
<tr>
<td>S5</td>
<td>-23</td>
<td>-22 (126547)</td>
<td>-23,-23 (245943, 532787)</td>
<td>-23,-23</td>
<td>-23</td>
</tr>
<tr>
<td>S6</td>
<td>-21</td>
<td>-21 (592887)</td>
<td>-21,-21 (365222, 691989)</td>
<td>-21,-21</td>
<td>-21</td>
</tr>
<tr>
<td>S7</td>
<td>-36</td>
<td>-34 (208781)</td>
<td>-35,-33.57</td>
<td>-35,-34.79</td>
<td>-36</td>
</tr>
<tr>
<td>S8</td>
<td>-42</td>
<td>-37 (187393)</td>
<td>-42,-42 (176313, 340917)</td>
<td>-42,-41.87</td>
<td>-42</td>
</tr>
</tbody>
</table>

Optimal configurations for three of the sequences: S5, S6 and S8

**S5 - 23**

**S6 - 21**

**S8 - 42**
Protein folding

Unfolded

Folded
Cellular automata

von Neumann neighborhood in a cellular automaton in 2D

Conway’s “Game of Life”
Protein folding with “Neural Cellular Automata”
Protein folding. Example in 2D with the HP model
Protein folding. Example in 3D with the HP model
Evolving “Neural Cellular Automata” with Differential Evolution

Genotype
Connection weights $w_{ji}$

Fitness function
Number of final HH contacts
Protein folding process modeling with the HP model

2D sequence – Optimum: 14 HH contacts

3D sequence – Optimum: 12 HH contacts
Protein folding process modeling with the off-lattice model of the Rosetta system

Protein conformation representation with the dihedral angles

Neural cellular automaton
Protein folding process modeling with the off-lattice model of the Rosetta system

Snapshots of the folding at the end of different temporal steps with protein sequence 1j4m. The last snapshot shows the native structure (green) and the final folded structure at the last step (blue).
Protein folding process modeling with the off-lattice model of the Rosetta system

Final folded conformation for protein 1j4m (14 amino acids) (blue) and the corresponding native structure (green) when the energy term associated with secondary structure elements is taken into account and when the folding process begins with a partially folded conformation.

Final folded conformation for protein 1d5q (27 amino acids) (blue) and the corresponding native structure (green) when the folding process begins with a partially folded conformation.
Tumor growth modeling with cellular automata
Cancer hallmarks

1. SG. Growth even in the absence of normal “go” signals.
2. IGI. Growth despite antigrowth signals issued by neighboring cells.
3. EA. Evasion of programmed cell death (apoptosis)
4. EI. Effective immortality.
5. GI. Genetic Instability.
6. AG. Ability to stimulate blood vessel construction (angiogenesis)
7. MT. Power to invade other tissues and spread to other organs.

AC Cell growth simulation

Cell “genome”

Mitosis event in the future: between 5 and 10 instant in the future

Each gene (hallmark) is mutated with a \( \frac{1}{m} \) chance of mutation

Daughter cell

Normal cell mitosis

1 iteration = 2.6 hours
5000 iterations = 75 weeks
AC Cell growth simulation. Self-growth hallmark (SG)

If *Self-growth* hallmark acquired, there is no need of growth factors.
AC Cell growth simulation. Ignore growth inhibit hallmark (IGI)

No empty space in the neighborhood

If Ignore growth inhibit hallmark acquired, a neighbor is killed (with probability $l/g$) to make room for mitosis
AC Cell growth simulation. Evade apoptosis hallmark (EA)

If Evade apoptosis hallmark (EA) is ON, there is not apoptosis.

A cell with \( n \) hallmarks mutated has an extra \( n/e \) likelihood of dying each cell cycle.
AC Cell growth simulation. Effective immortality hallmark (EI)

If telomere reaches 0, the cell dies.

Even if telomere length = 0, if Effective immortality hallmark (EI) is ON, the cell can continue dividing.
AC Cell growth simulation. Genetic instability (GI)

If Genetic instability factor (GI) is ON

Increment the cell base mutation rate by a factor \(i\)

Push mitotic events in the future: between 5 and 10 instants in the future

Schedule future mitosis

Is possible the mitotic division?

No

Yes

Schedule future mitoses for both cells
Simulating cancer cell cultures: Multicellular spheroids

Microscopy image of a multicellular tumor spheroid, exhibiting an extensive branching system that rapidly expands into the surrounding extracellular matrix gel. These branches consist of multiple invasive cells. Guiot et al. Theoretical Biology and Medical Modelling 2007 4:4 doi:10.1186/1742-4682-4-4

Growth stages of 3D-cultured CT26 colon cancer spheroid. Five hundred suspension cancer cells were dispensed into each well of a 48-well culture tray. Trays were then inverted and incubated during 12 days. Spheroids were collected on days 3, 5, 7 and 12 and processed for cell counting, spheroid diameter determination and immunohistochemical detection of Ki67-expressing cells. Scale bar: 100 μm.

Some simulations and results
Dependence on hallmark parameters

Evolution through time iterations of the number of healthy cells (continuous lines) and cancer cells (dashed lines) for different base mutation rates (1/m).
Simulation run with default parameters and $m=100$

Grid size = 125000, mutation rate = 0.01
Simulation run with parameters that facilitates cancer appearance

Grid size=125000, mutation rate=$10^{-5}$
Simulation run with parameters that facilitates cancer appearance

Snapshots of the cellular system at different time steps

Grid size=$10^6$, mutation rate=$10^{-5}$
Relevance of hallmarks

Effect of elimination of individual cancer hallmarks. Default parameters.
Behavior transitions

Effect of killing cancer cells during tumor growth for different killing probabilities and using four parameter sets.

Killing only outer cancer cells
Cancer Stem Cells

CSCs can divide symmetrically or asymmetrically to produce Differentiated Cancer Cells (DCCs) with limited proliferative capability.
Simulating Cancer Stem Cells

Introduction of Cancer Stem Cells (CSCs) in the multicellular system evolution. Standard parameters and $g=5$. 

Example with high invasion potential.
Simulating Cancer Stem Cells

Snapshots of 2D central sections

Gray: Healthy cells, Blue: DCCs, Red-enlarged size: CSCs.
Simulating Cancer Stem Cells

Example with high mutation rate

Standard parameters and m=1000.
Simulating Cancer Stem Cells

Snapshots of 2D central sections
Gray: Healthy cells, Blue: DCCs, Red-enlarged size: CSCs.
FOLFOX4 treatment for colon cancer
Treatment options

Treatment applied every 600 time iterations killing the 100% of the DCCs in such iterations.

Treatment applied every 600 time iterations killing the 5% of the DCCs during the next 150 iterations.

2D snapshots of the central part of the grid at given time iterations (Colors: Gray - healthy cells, Blue - DCCs, Red-enlarged size - CSCs).
Continuous treatments that kill 1%, 2% and 3% of DCCs are applied in every time iteration.
Treatment options

Accumulative treatment intensity across iterations

Accumulative treatment intensity

- killing 1% of DCCs (Fig 2)
- killing 2% of DCCs (Fig 2)
- killing 3% of DCCs (Fig 2)
- killing 5% of DCCs (Fig 4)
- killing 100% of DCCs (Fig 3)
In this application: F:0.9
CR:0.9
X₁ selected with tournament

Encoded parameters: treatment intensity, duration and period.
50 generations, grid size of 64000 sites.
Hallmark parameters: g = 5, m = 1000, others - standard values
Fitness: DCCs after a given number of iterations (100) once the treatment is stopped +
Intensity used during the treatment application.
Evolved treatment strategies

Treatment intensity and duration optimized. Treatments begin in time iteration $t=600$. The best evolved treatment kills 38% of DCCs in the next 18 time iterations. A high-intensity treatment that kills the 100% of DCCs only at $t=600$ is included for comparison.
Evolved treatment strategies

Evolution of DCCs

Treatment period, intensity and duration optimized. Treatments do not begin before time iteration $t=600$. The best evolved treatment kills 49% of DCCs during 8 time iterations and a period of 346 iterations (treatment A). A high-intensity treatment that kills the 100% of DCCs every 600 iterations is included for comparison (treatment B).
Conclusions

Our focus was on the dependences of the first phases of cancer growth on the hallmark parameters:

- In a CSC context, the model predicts and is in agreement with the clinical observations describing *increased growth speed and enhanced invasion in the relapsing malignancy.*
- Using EC: Treatments should be maintained during very few days, avoiding high intensities and, consequently, with longer periods than the ones used in standard treatments. The aim is to **make more difficult CSC proliferation** and consequently their differentiation to minimize their future effect on a possible tumor regrowth.

The hallmark implications for cell population dynamics are difficult to foresee without a simulation model.

The modeling can lead to a better understanding and characterization of the underlying biological processes involved.
References


