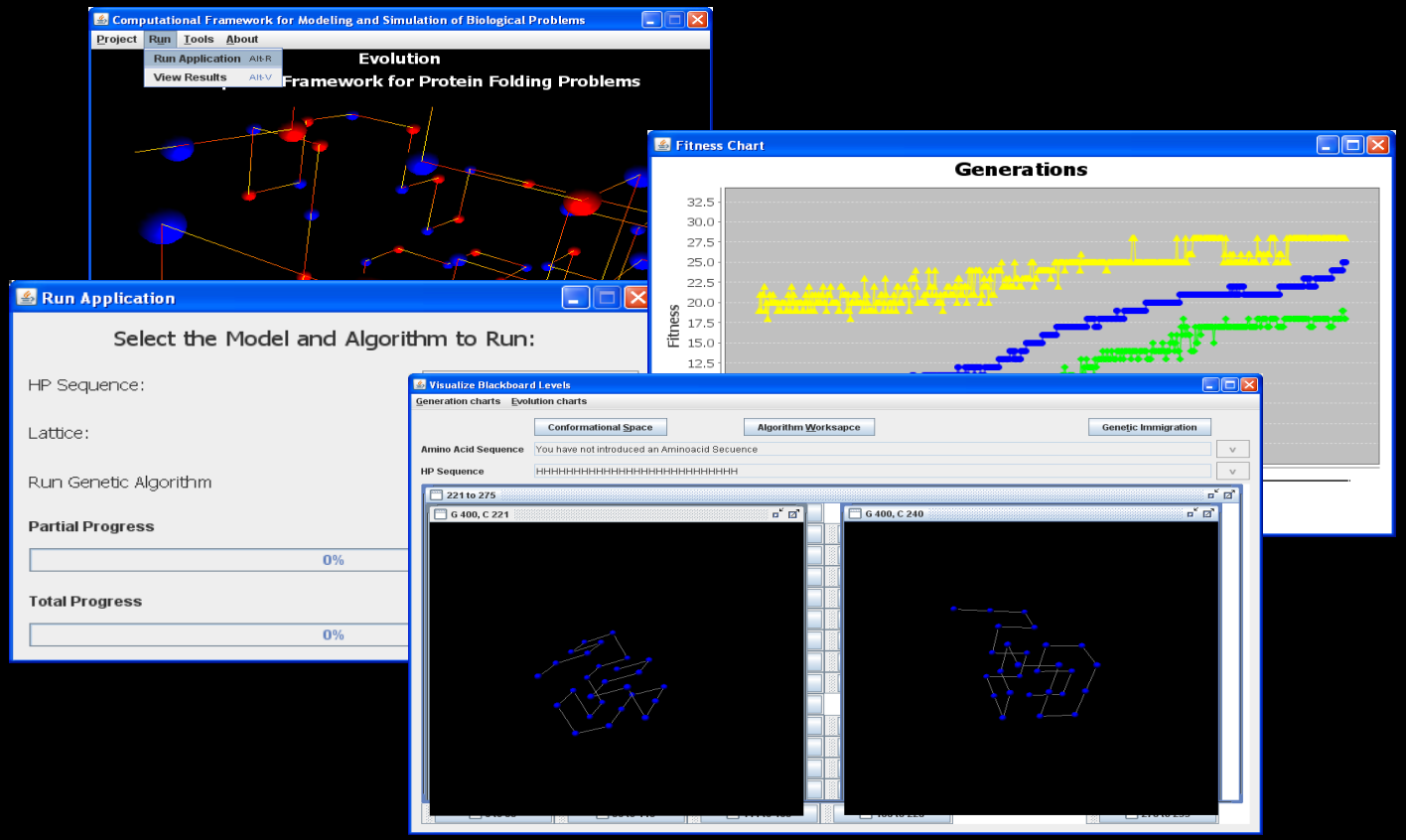
Exploring dimensionality, systematic mutations and number of contacts in simplistic HP *ab-initio* protein folding using a blackboard-based agent platform

González-Pérez, P.P., Rojo-Domínguez, A., Beltrán, H., and Sánchez, M.

Universidad Autónoma Metropolitana, México



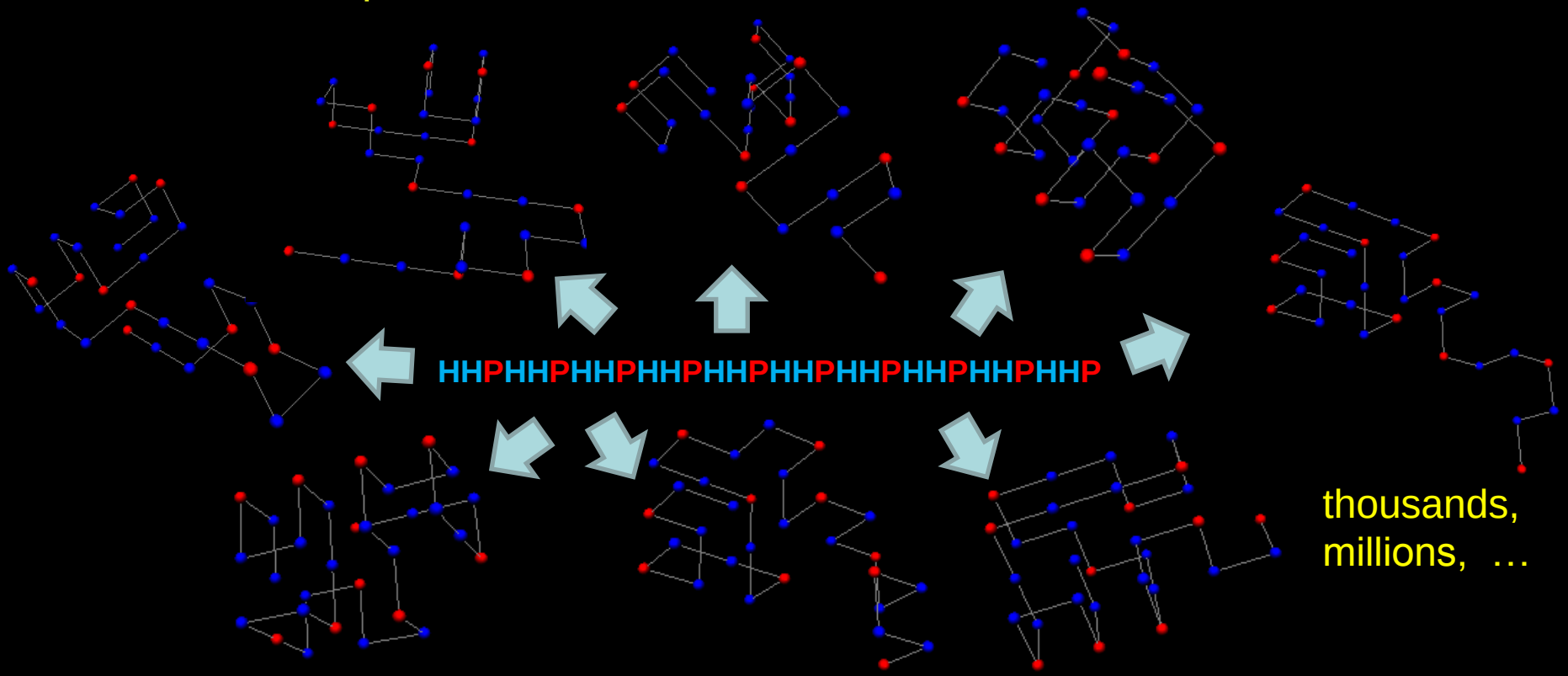
Complexity in biological systems (I)

Biological systems have the characteristic of having a high complexity which leads to two problems:

- On the one hand, they have a huge number of degrees of freedom, and moreover,
- the evaluation function is extremely difficult to calculate accurately.

Complexity in biological systems (II)

- The first problem means that instead of choosing between a few alternatives, we have to choose between thousands, millions, billions or even powers of ten that can be higher than 10^{100}, which means that even the most powerful computer in the world will require astronomical times to complete the task.



Complexity in biological systems (III)

- The second problem, which refers to the evaluation function, has been tackled with the use of highly simplified models of biological phenomena, which has been able to significantly reduce the computational cost of evaluating each alternative.

Simplified models for protein folding

$$F = f(e)$$

$$F = f(e) + g(R_{gyration}) + h(D_{\max})$$

Examples of complex systems in Structural Biology

Folding

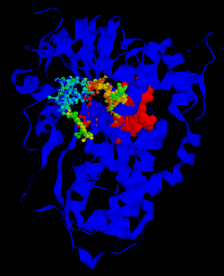
meaiakydfk ataddelsfk
... qtgmfprnyv tpvnrnv

Inverse folding

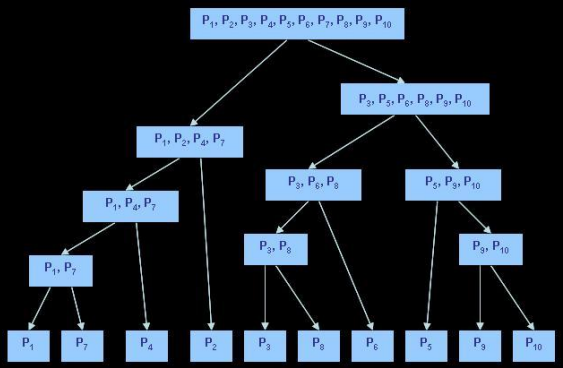
mteyk…vg … iqliqnhfvd
... kskakcvvm

mte… aggvgksalt
iqliqnhfvd ... ksk.. vvm

Docking



Clustering



Structure-activity and structure-property relationships

$$activity = f(structure)$$
$$activity = c_0 + c_1 MD_1 + c_2 MD_2 + \cdots + c_n MD_n$$
$$property = f(structure)$$
$$property = c_0 + c_1 MD_1 + c_2 MD_2 + \cdots + c_n MD_n$$

Sequence alignment

The protein folding problem

Protein folding:

- The mechanism by which a protein adopts its biologically active conformation after biosynthesis.
- *The protein folding problem is essentially a search for the biologically active (functional) conformation of a protein (the so-called "native state"), for a given sequence of amino acid residues.*

meaiakydfk ataddelsfk ... qtgmfprnyv tpvnrnv

Grb2

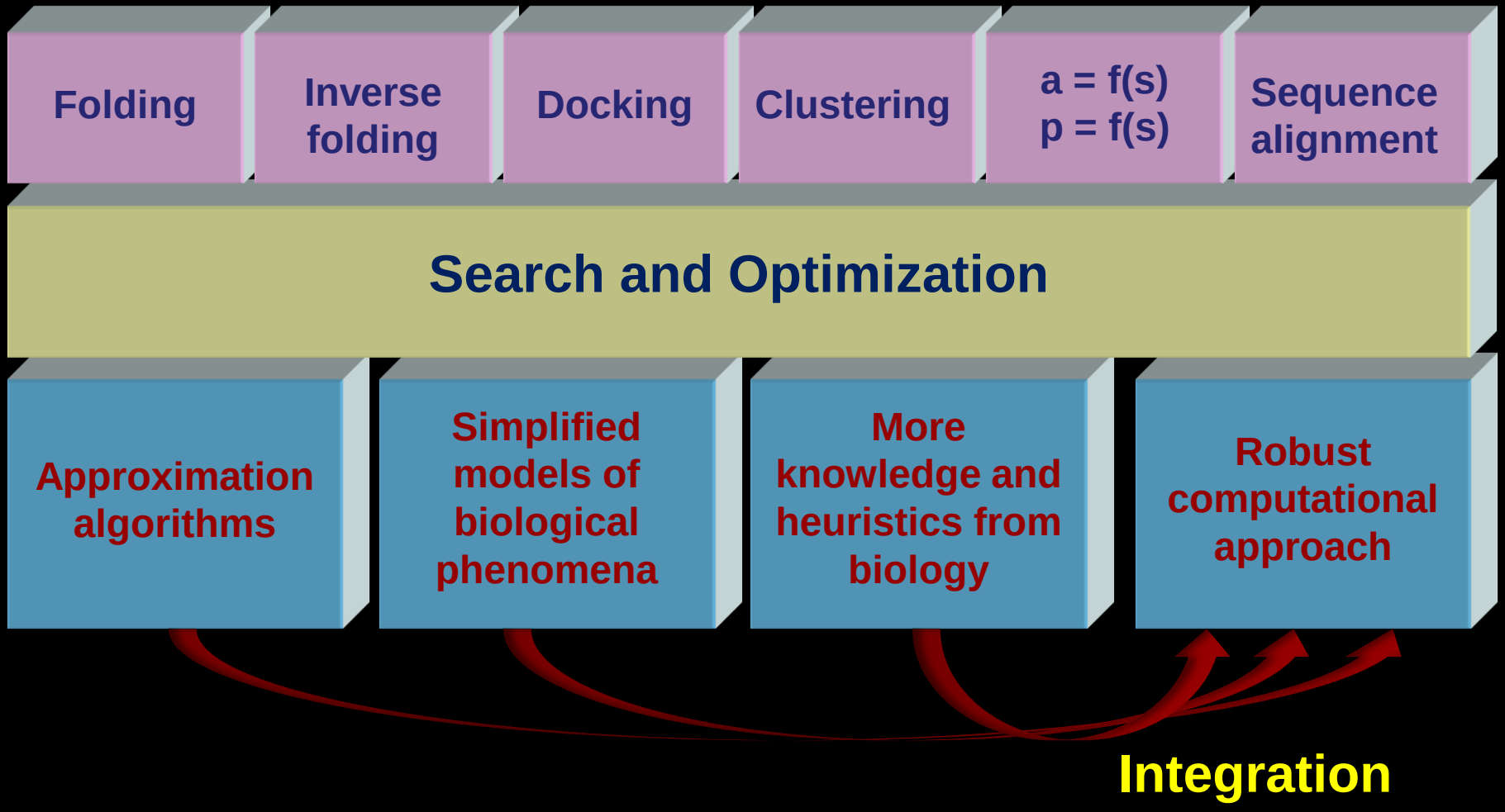
mddsevesta silasvkeqe drsgdlgdme ... plkgaplmqk i

p120

mteyklvvvg aggvgksalt iqliqnhfvd ... kskakcvvm

Ras

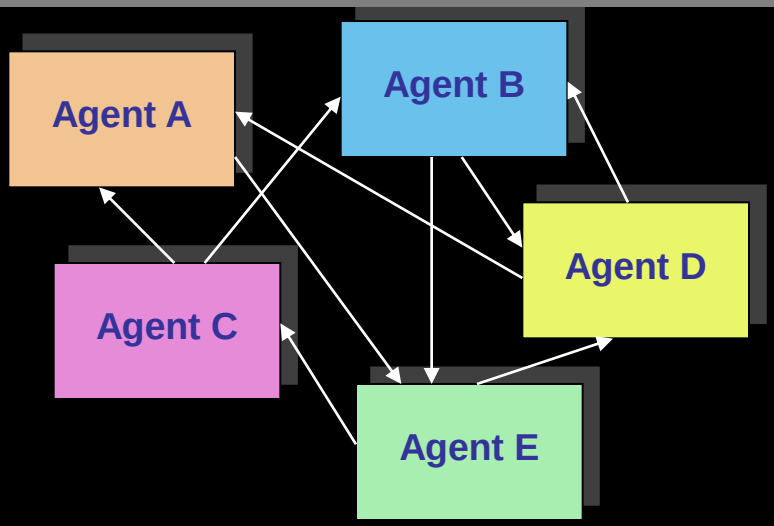
What is common among these biological problems?



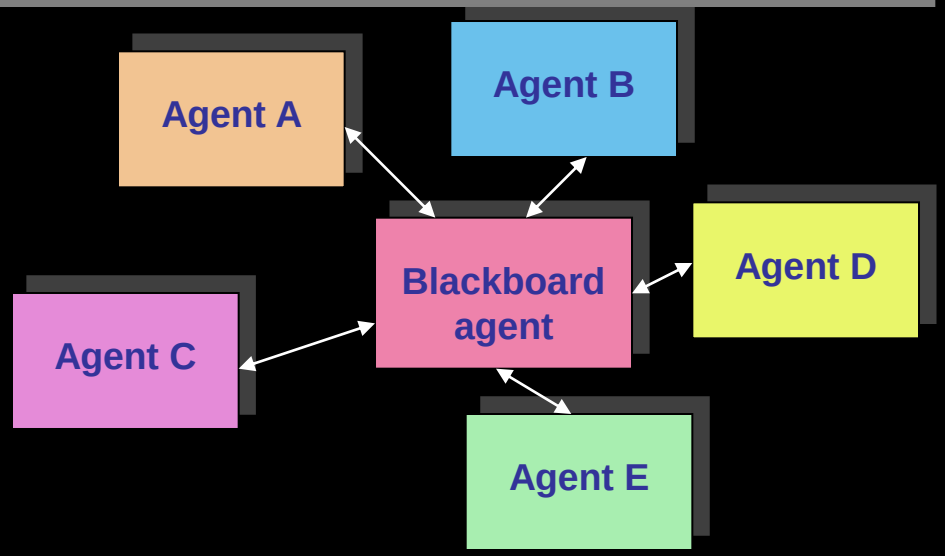
The purpose of this work

- The integration of the most robust computational techniques for generation, searching, optimization and visualization of molecular structures in a single adaptable computational framework
- Such a computational framework must allow the user to study and explore complex problems, such as those from structural biology, by providing the means to experiment with the system and to develop functionalities and algorithms to adapt the construct to the singularities of the problem.

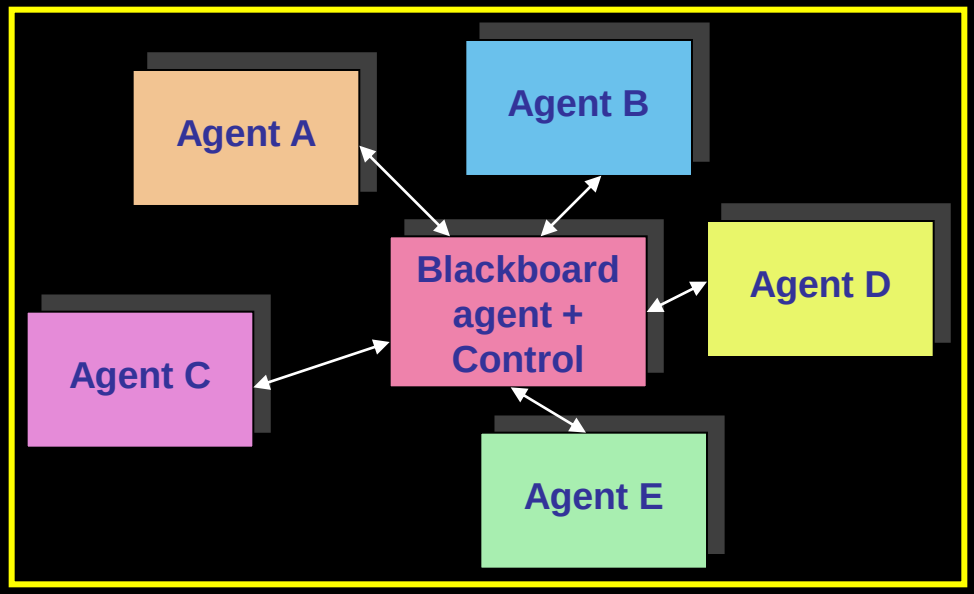
The computational approach: Collaborating agents (Corkill, 2003)



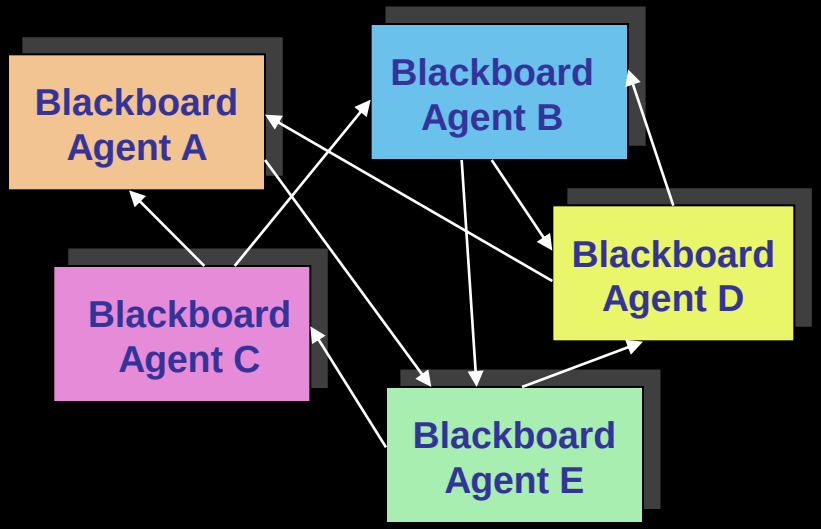
A: Directly interacting



B: With a Blackboard agent



C: With a Blackboard & control agent



D: Full-fledged Blackboard agents

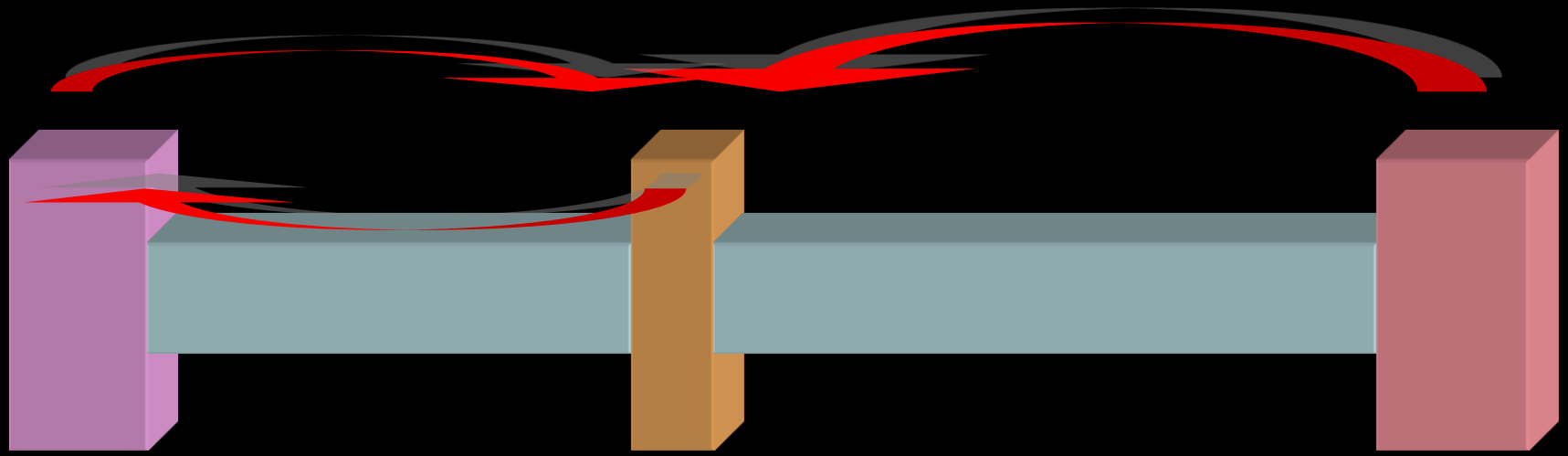
Between bioinformatics applications and multi-agent systems theory

Our contribution lies here

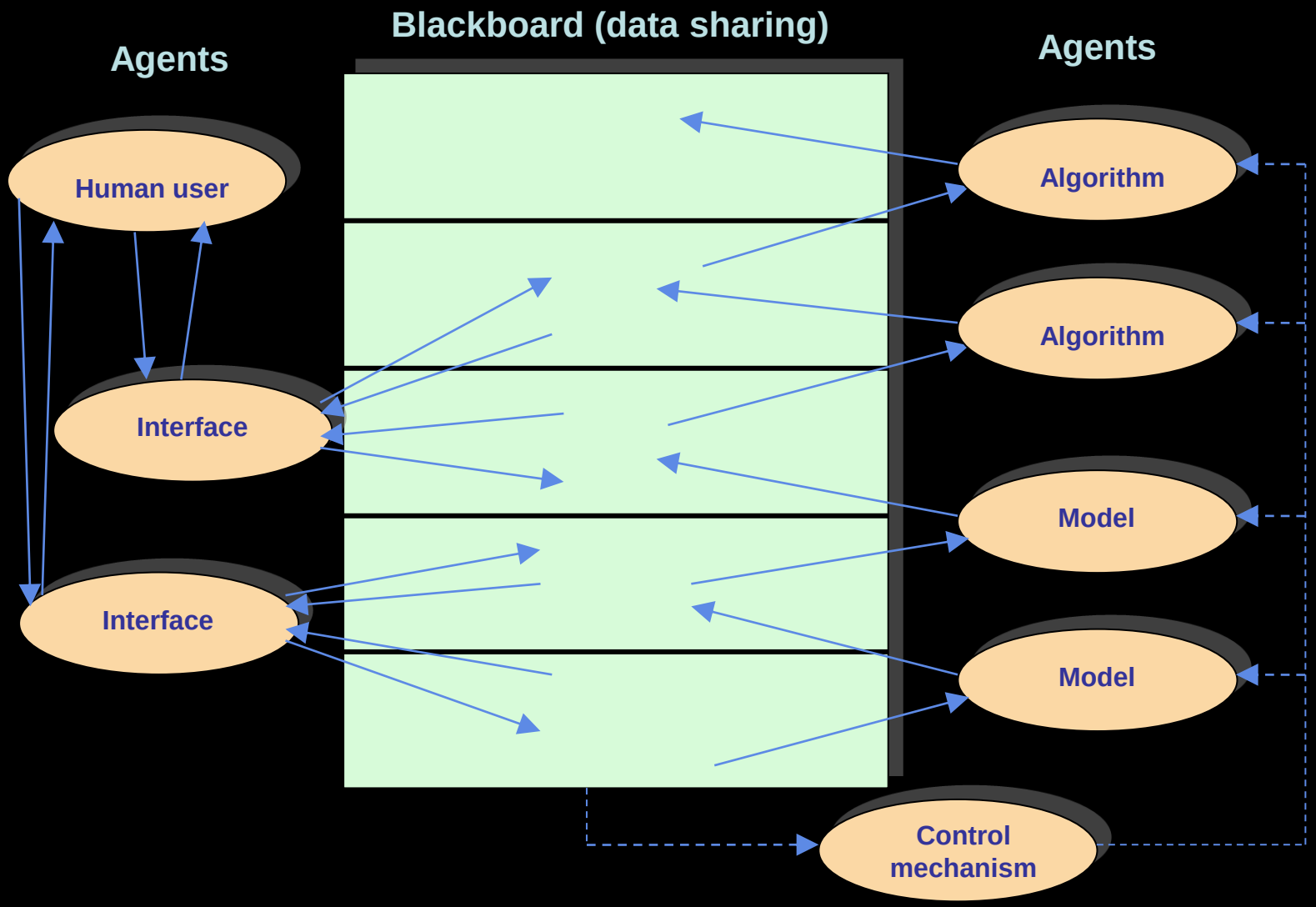
Unsolved problems in structural biology

Development of bioinformatics applications

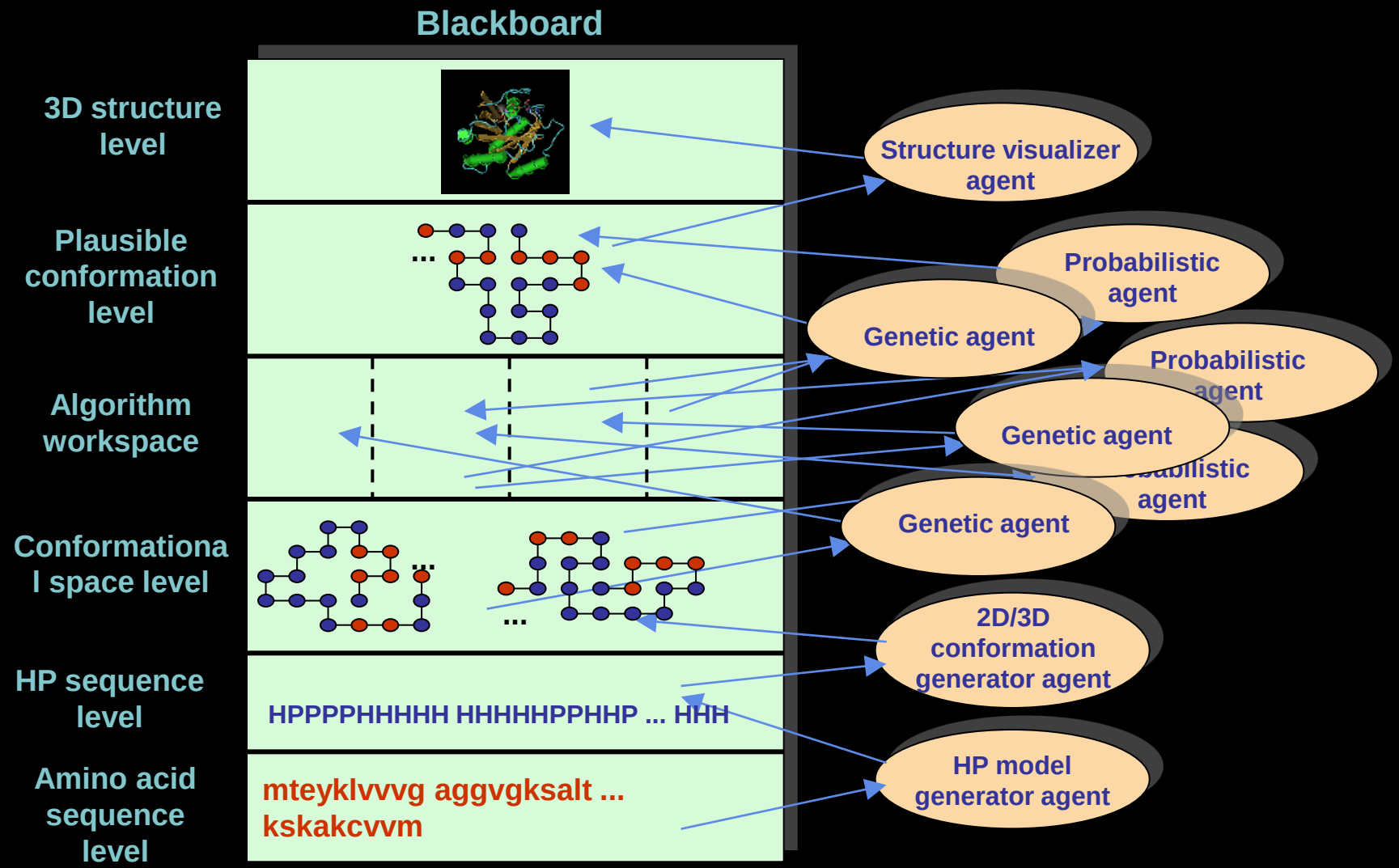
Blackboard and multi-agent systems theory



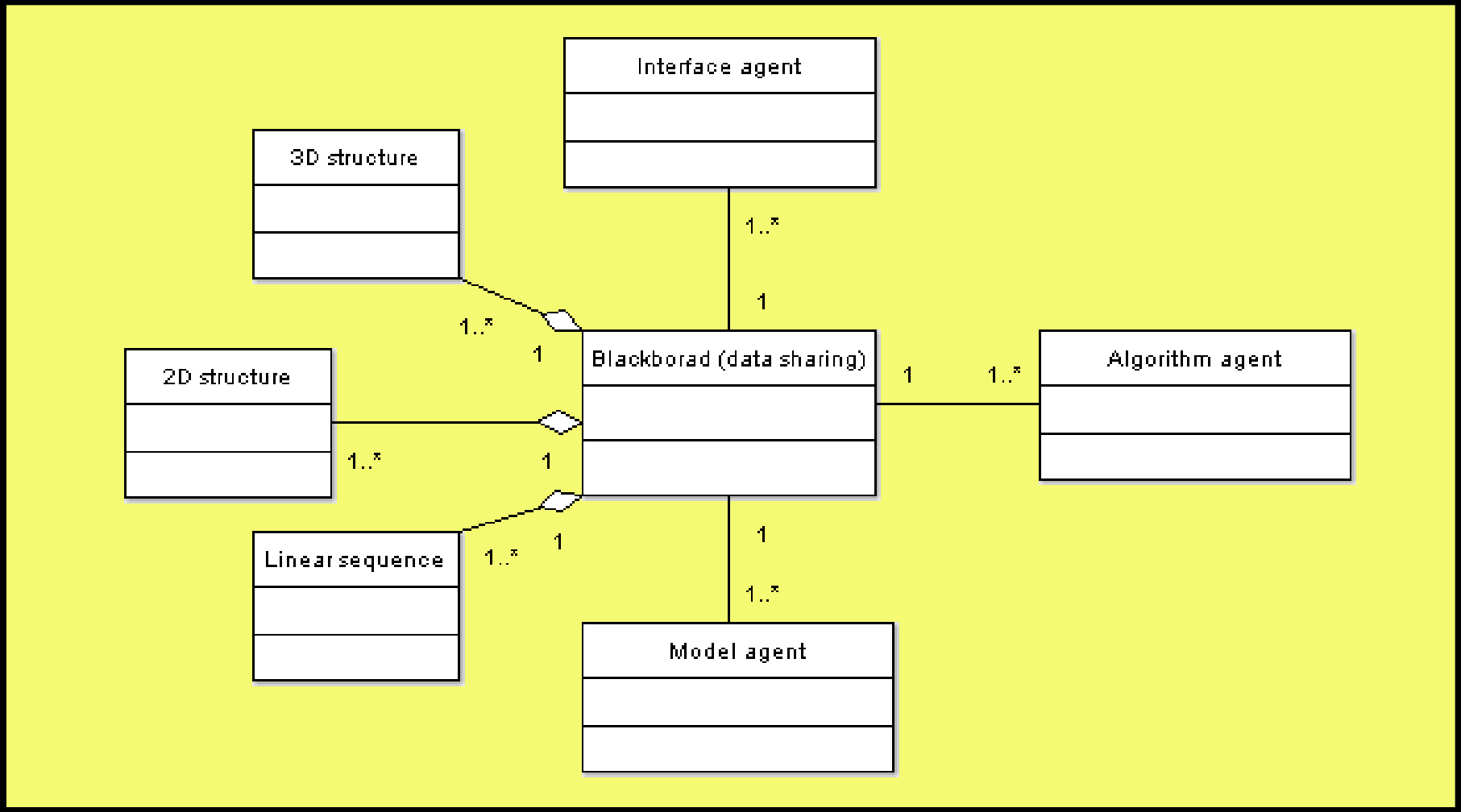
Blackboard-based multiagent platform: the architecture of bioinformatics framework



Blackboard-based multi-agent architecture: the semantics for protein folding problem

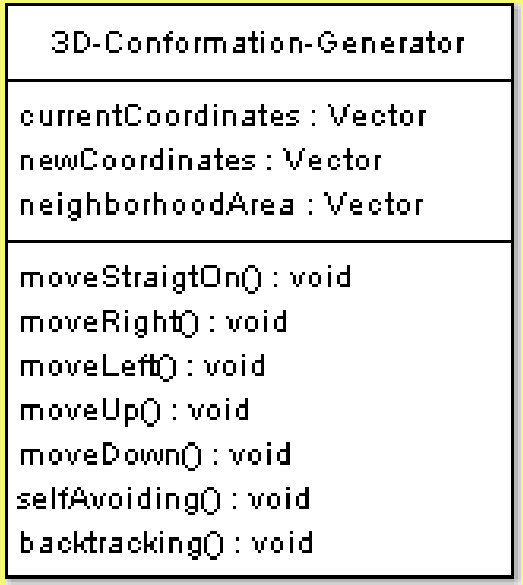


Blackboard-based multi-agent architecture: the architectural pattern

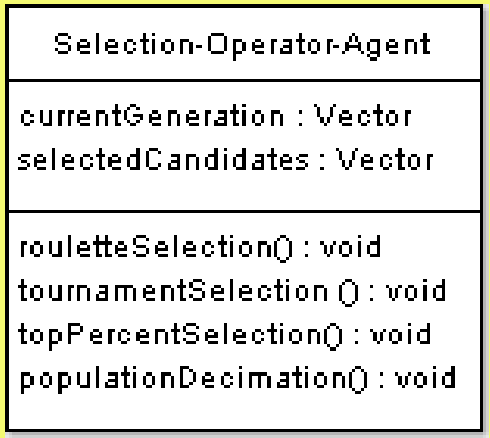


Blackboard-based multi-agent architecture: the agents

Model agent

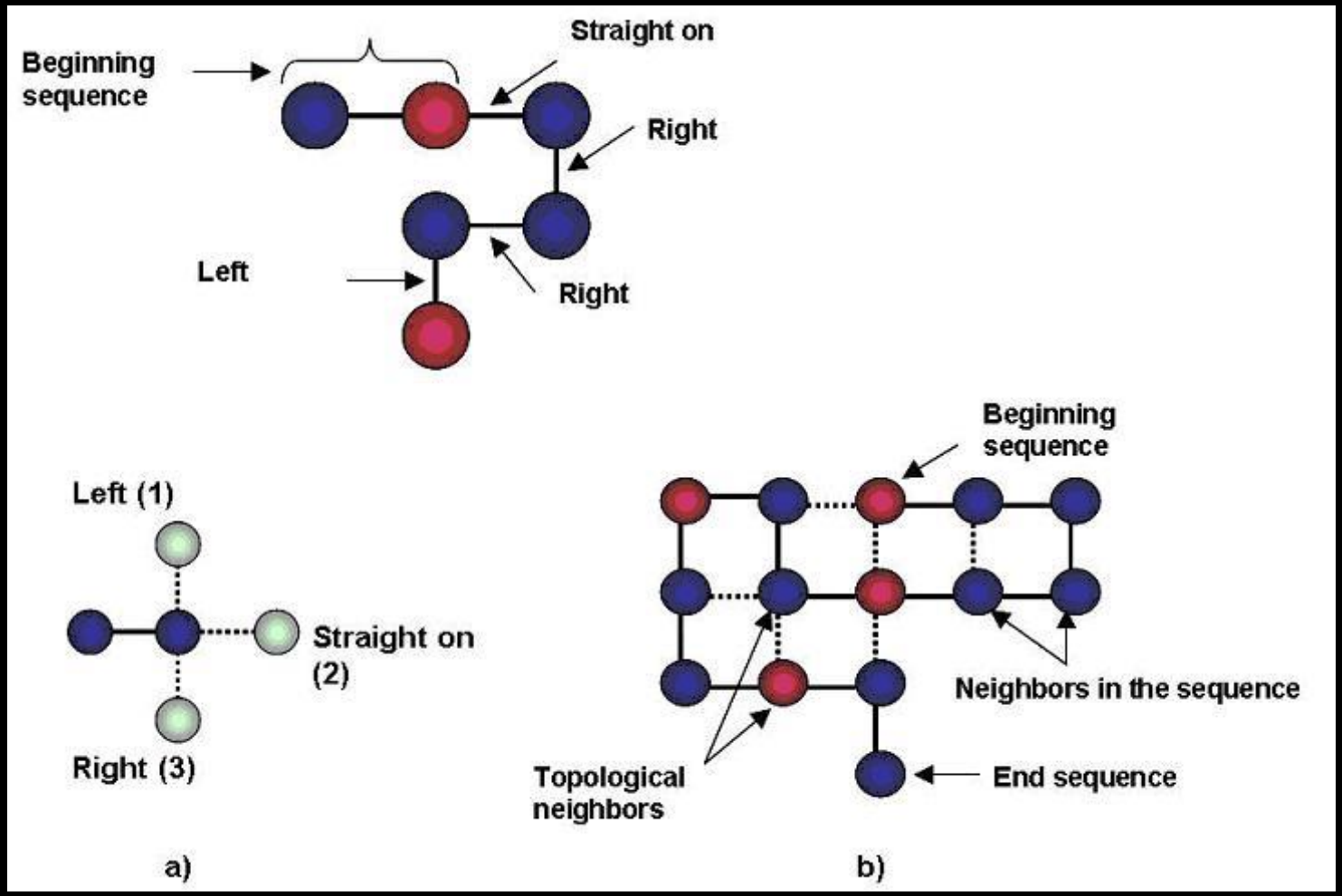


Algorithm agent



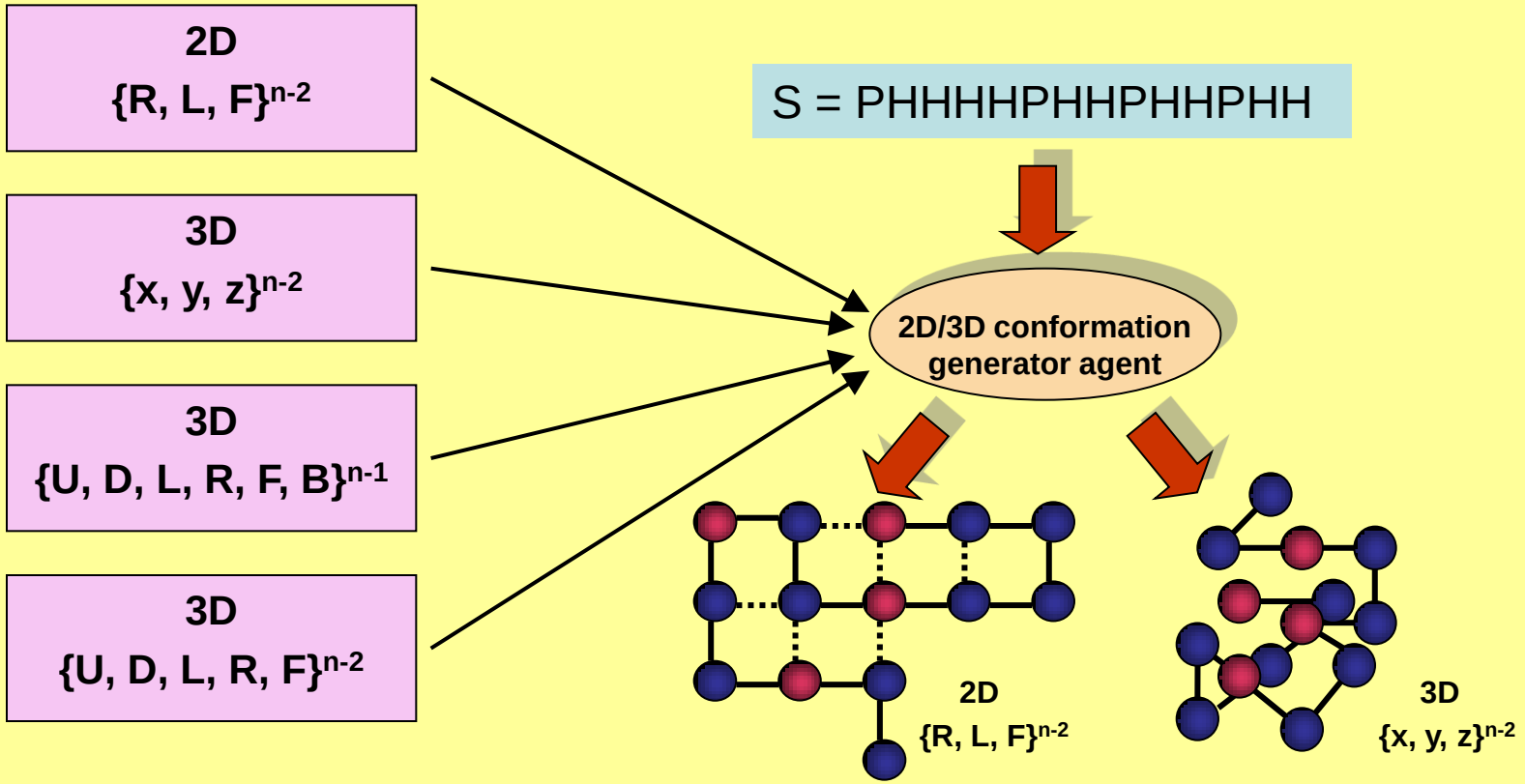
Blackboard-based multi-agent architecture: models and algorithms encapsulated within agents

A 2D space, with a square lattice



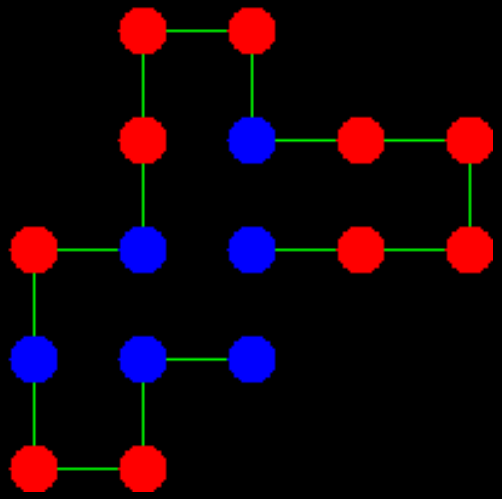
Blackboard-based multi-agent architecture: models and algorithms encapsulated within agents

Local coordinate systems

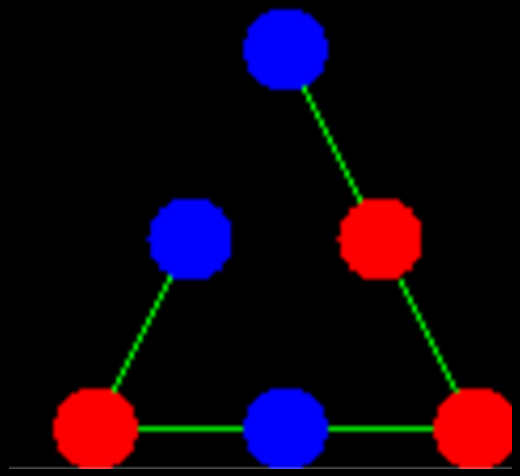


Blackboard-based multi-agent architecture: models and algorithms encapsulated within agents

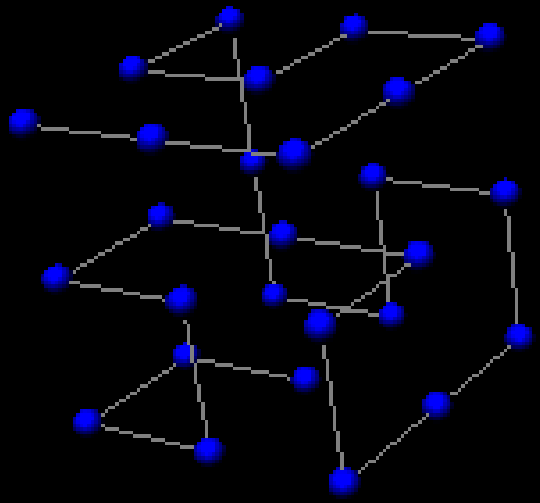
Different types of discrete spaces can be generated in Evolution



2D Squared

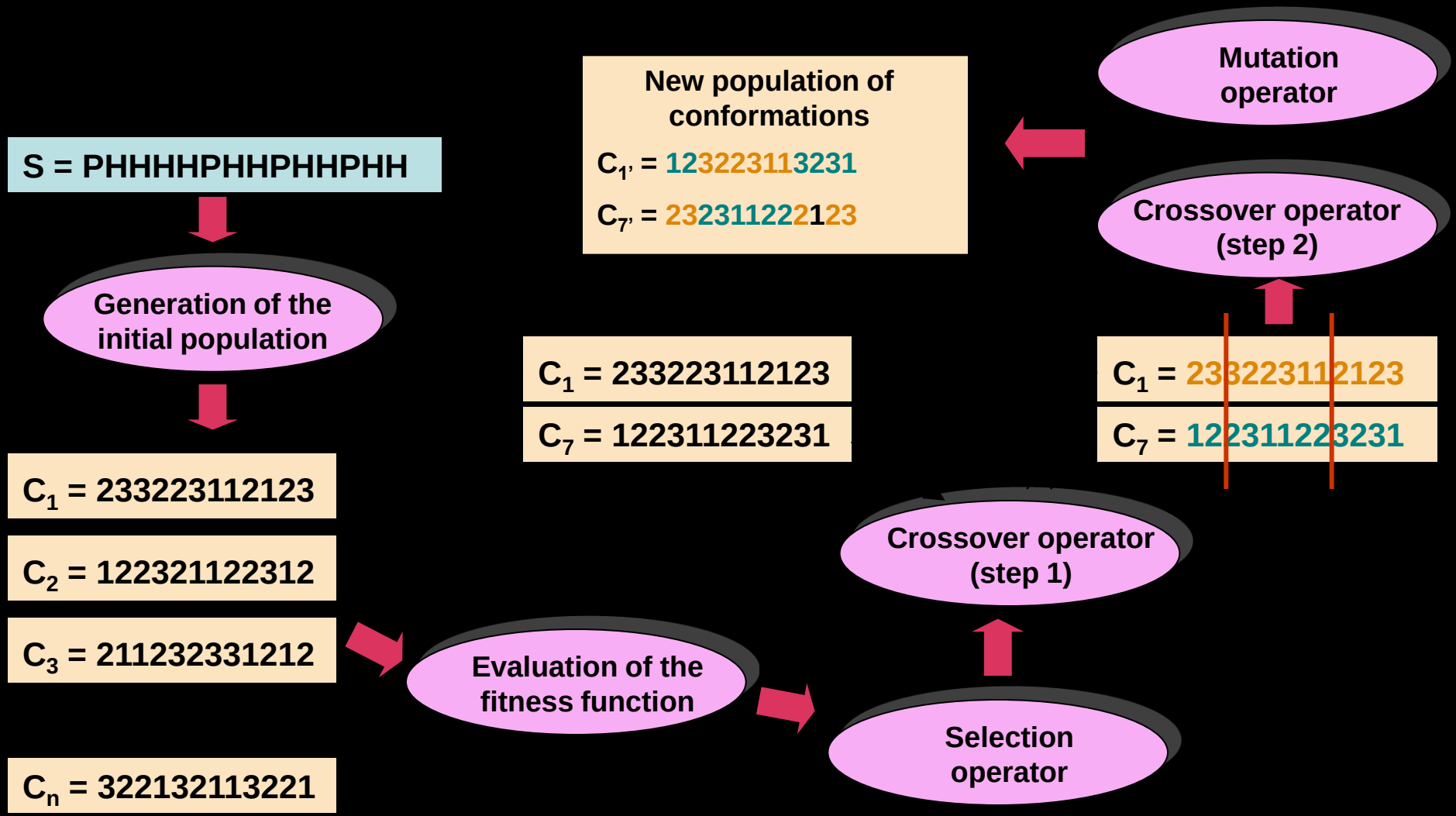


2D Triangular



3D Cubic

Blackboard-based multi-agent architecture: models and algorithms encapsulated within agents



Blackboard-based multi-agent architecture: models and algorithms encapsulated within agents

Model 1: the traditional model

$$F = f(e) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} e_{ij} \Delta_{ij}$$

$$e_{HH} = -1.0 \quad e_{HP} = 0.0 \; e_{PP} = 0.0$$

$$\Delta_{ij} = \begin{cases} 1 \; if \;\; i \; and \;\; j \; are \; topo\log ical, \; but \; not \; sequence \; neighbours \\ 0 \; otherwise \end{cases}$$

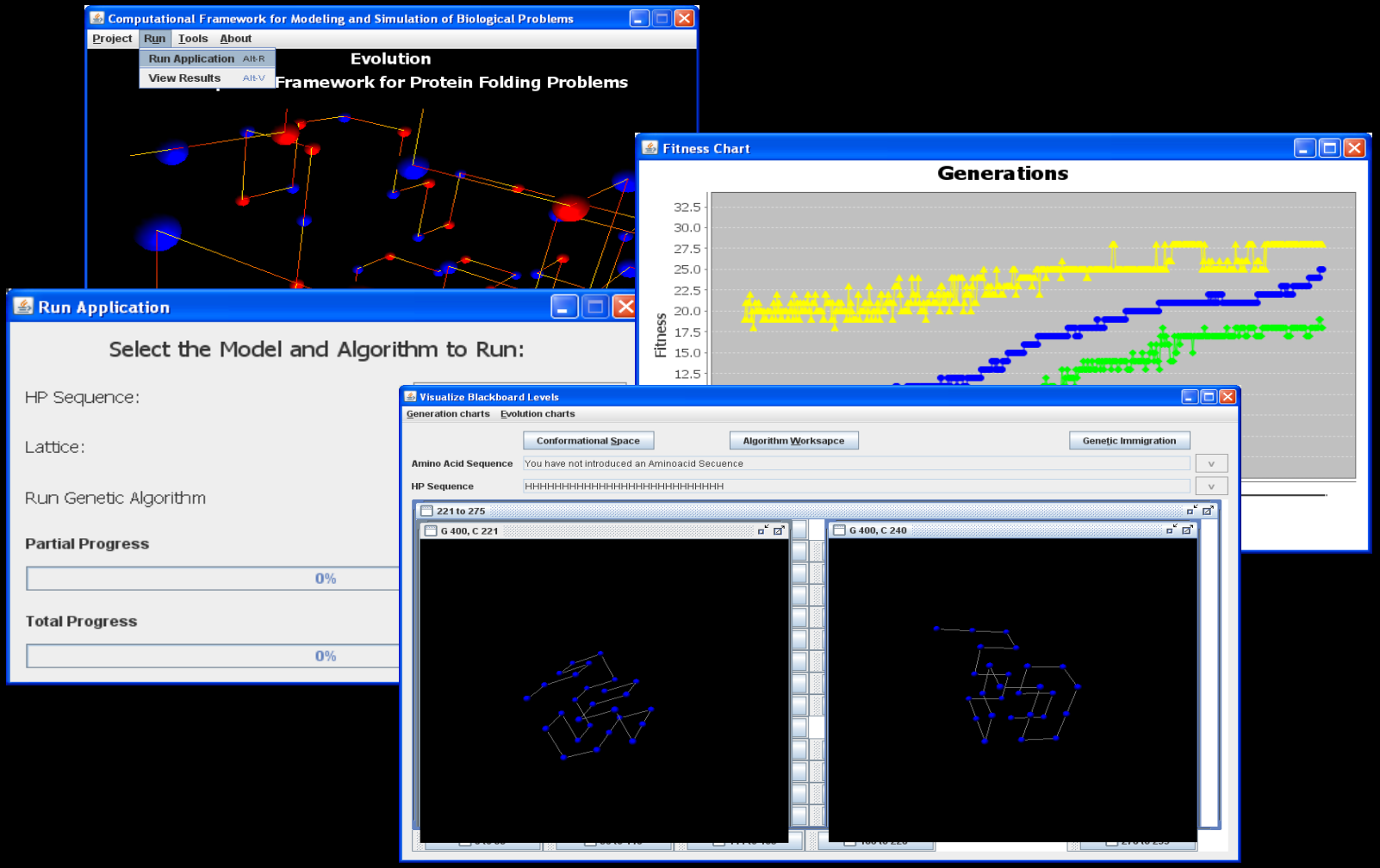
Model 2: there is a reward for PP bonds

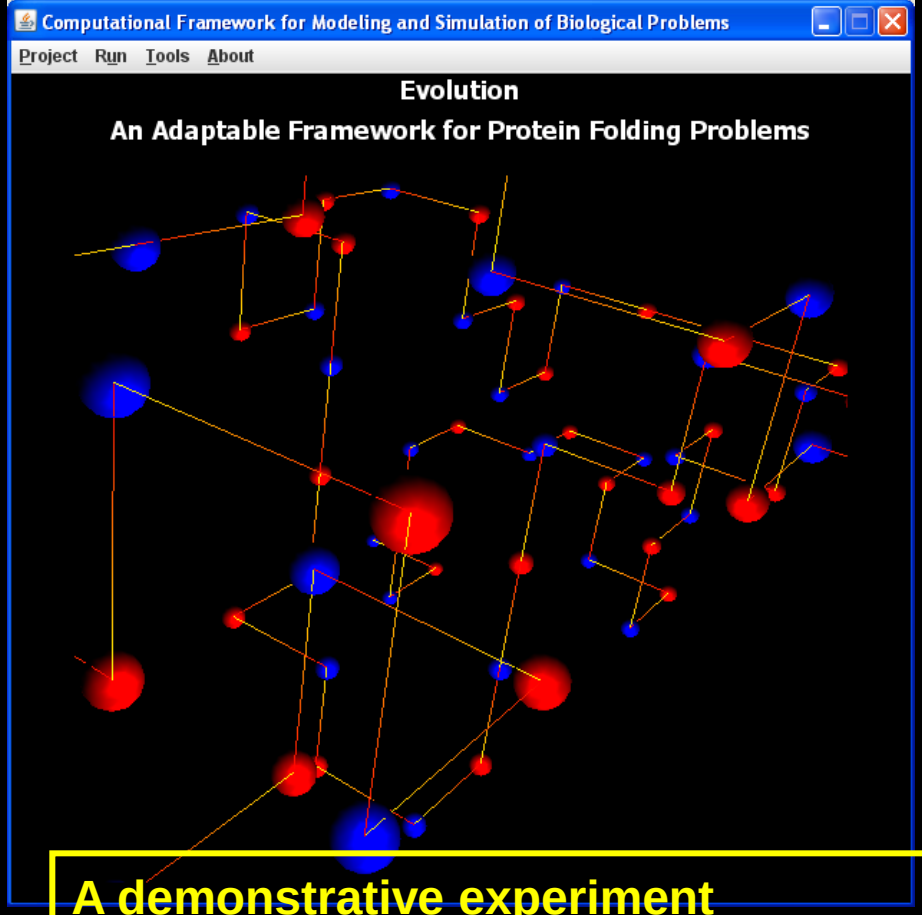
$$F = f(e) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} e_{ij} \Delta_{ij}$$

$$e_{HH} = -1.0 \quad e_{HP} = 0.0 \quad e_{PP} = -0.5 \; or \; ???$$

$$\Delta_{ij} = \begin{cases} 1 \; if \;\; i \; and \;\; j \; are \; topo\log ical, \; but \; not \; sequence \; neighbours \\ 0 \; otherwise \end{cases}$$

Blackboard-based multi-agent architecture: the virtual laboratory



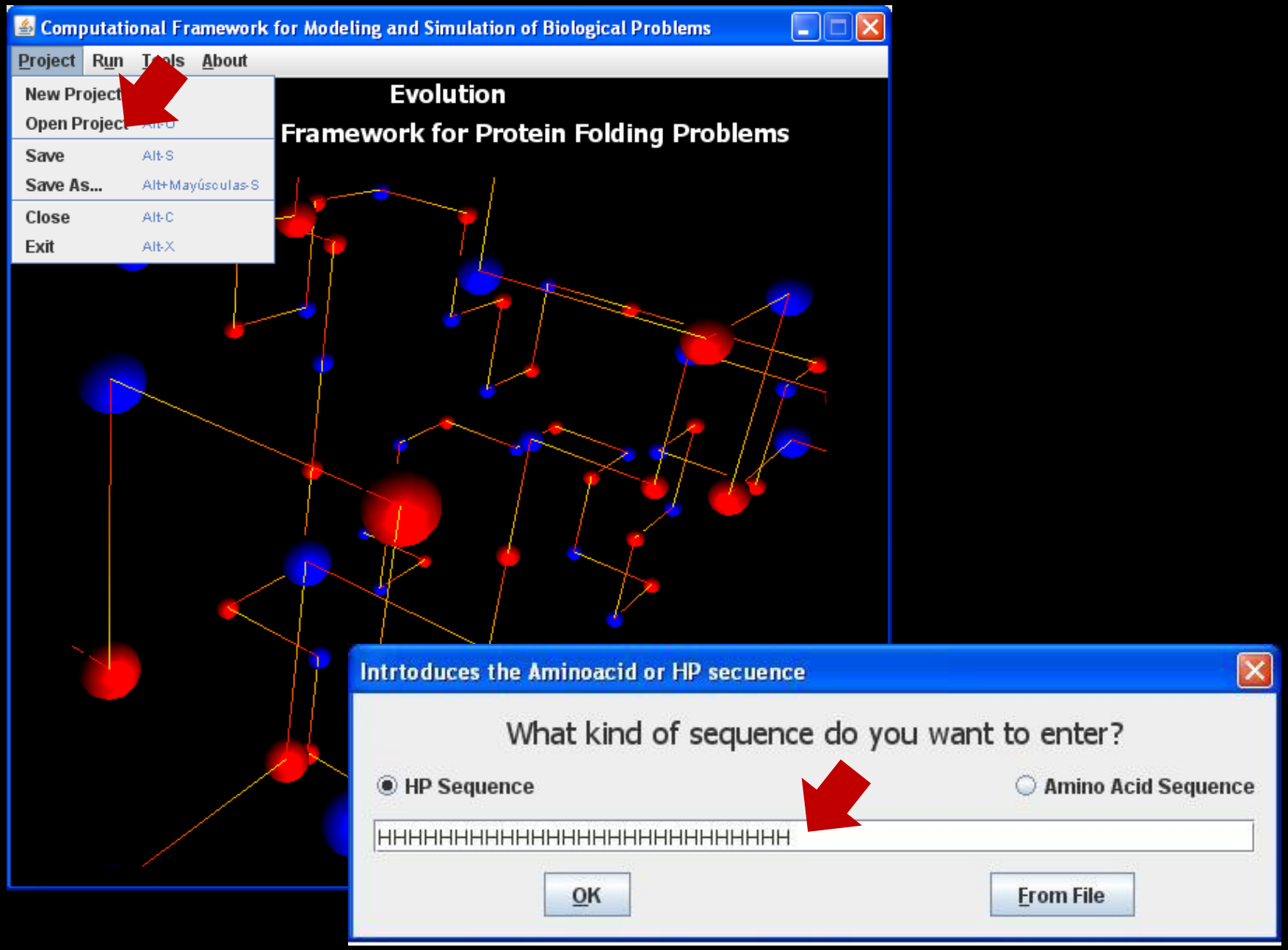


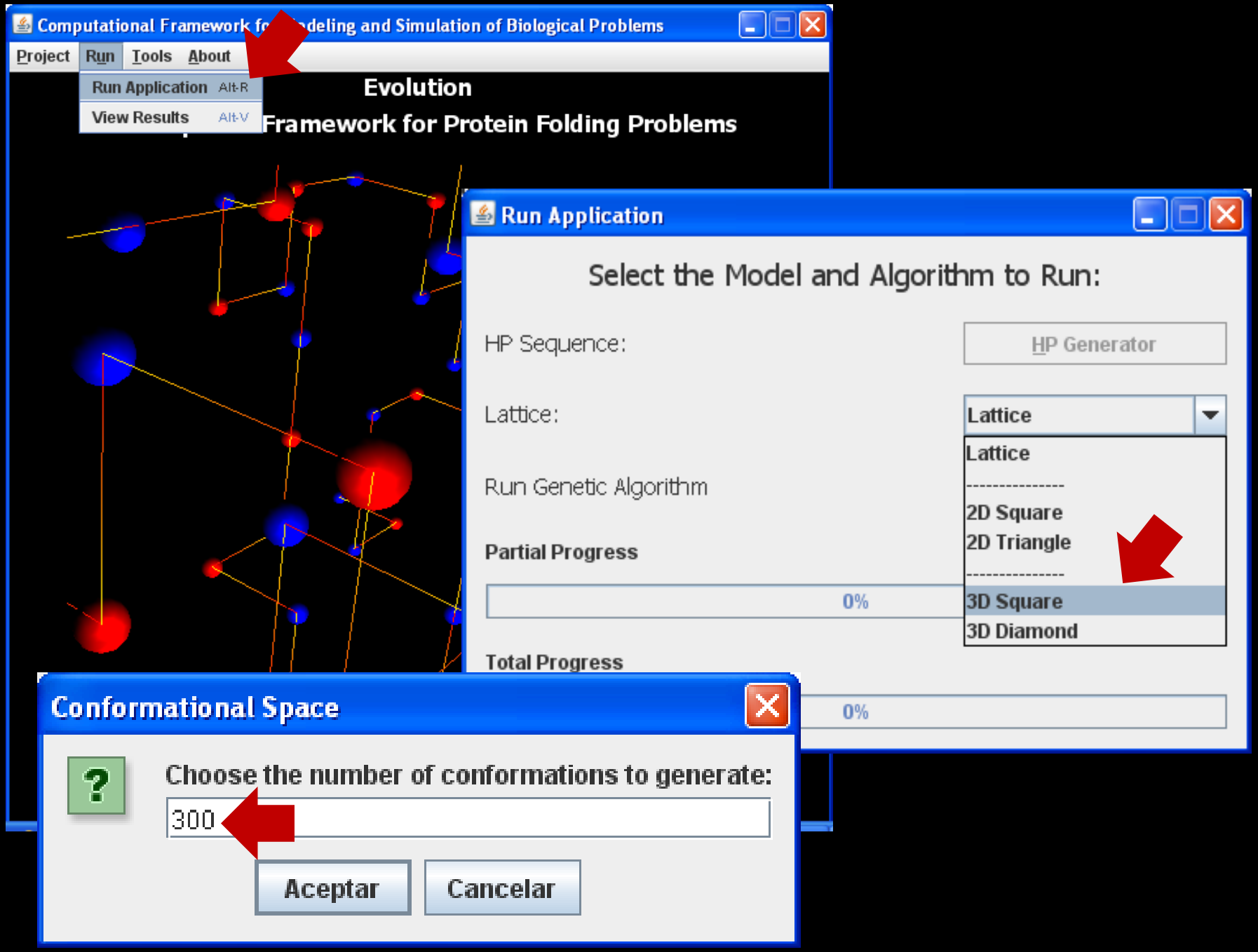
A demonstrative experiment

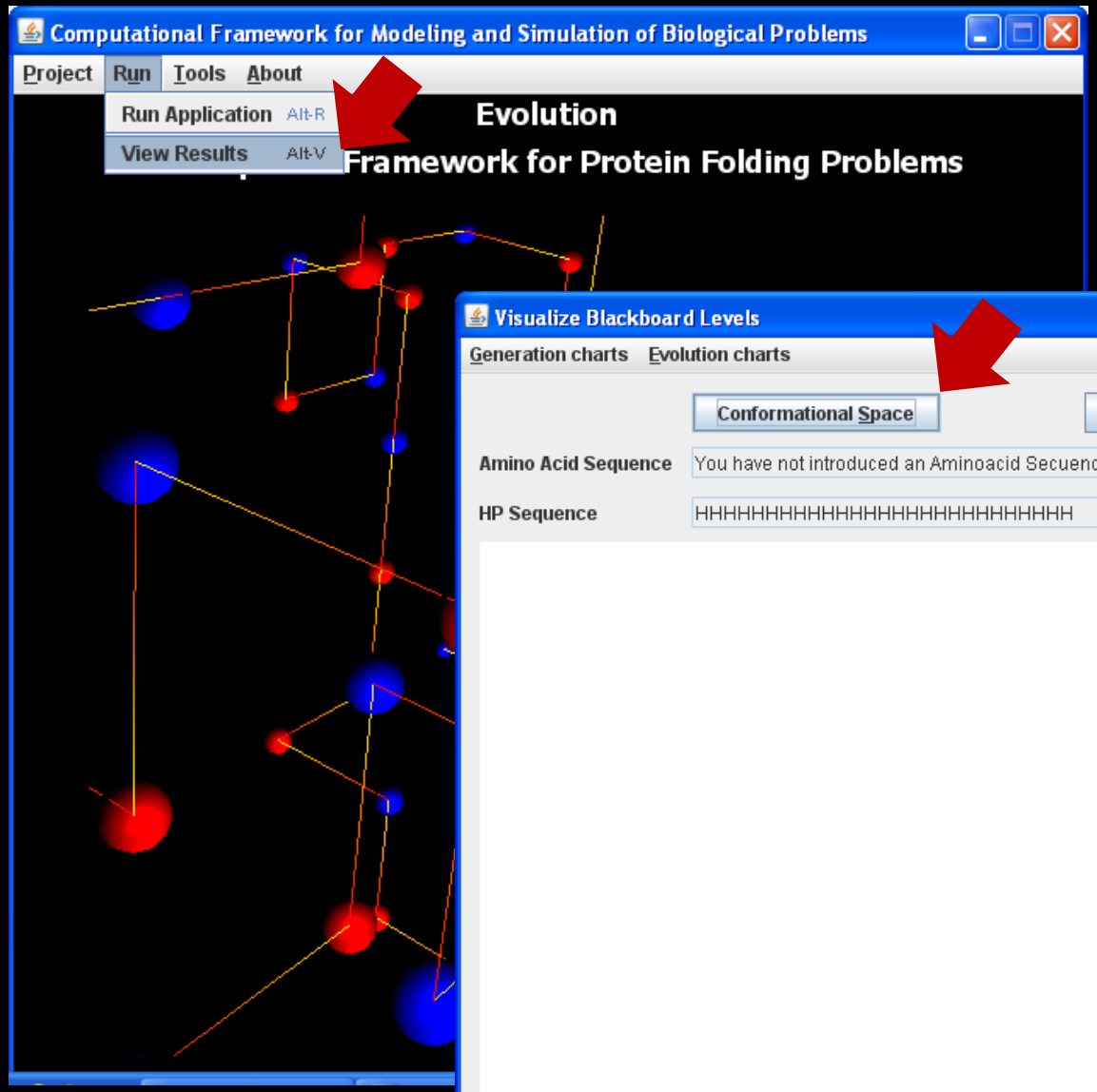
Sequence: HHHHHHHHHHHHHHHHHHHHHHHHHH

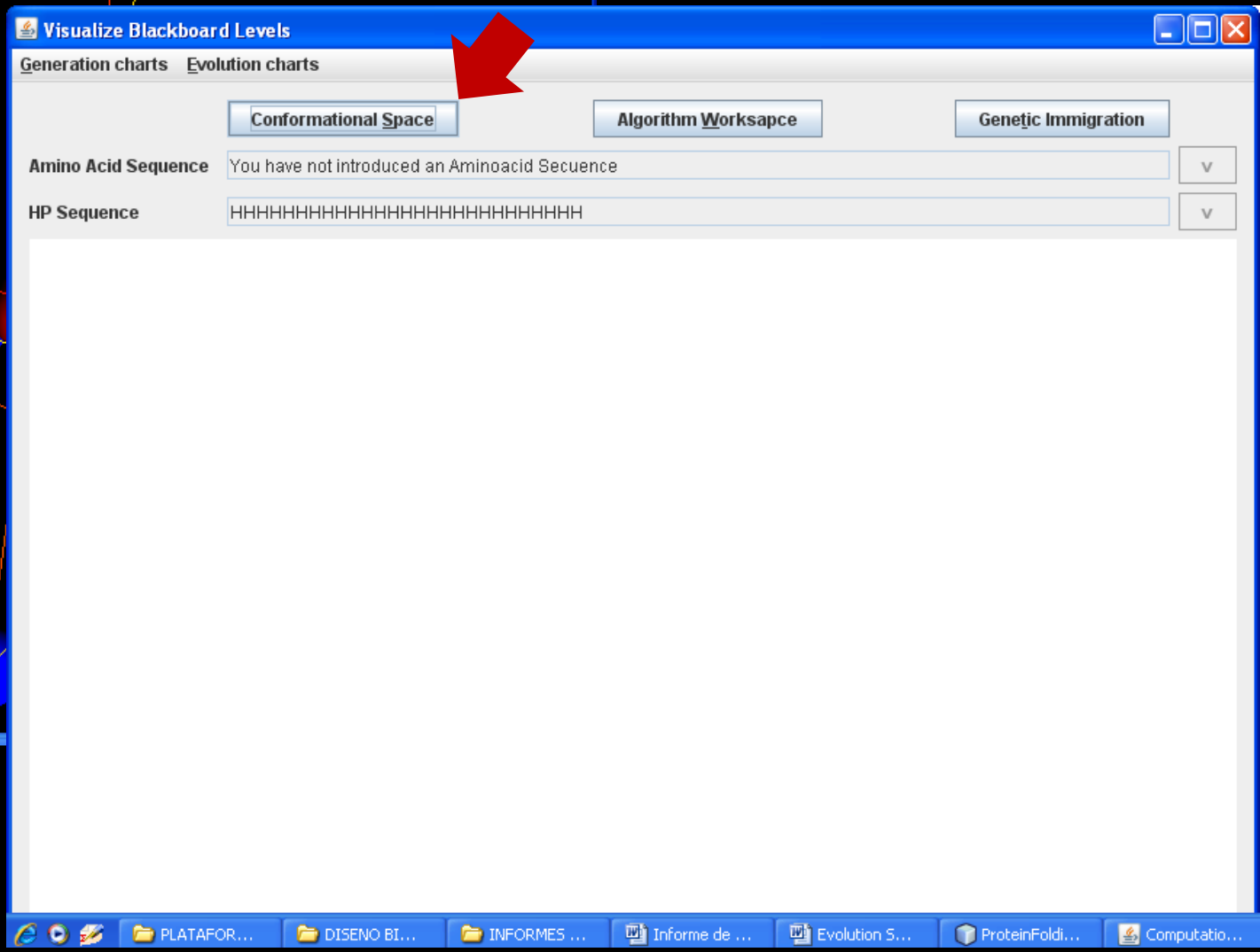
Parameters of the experiment:

- **Number of conformations per generation:** 300
- **Number of generations to be run by the evolutionary algorithm:** 500
- **Genetic operators:** *Roulette selection* method, *Single point crossover* and *Simple mutation.*
- **Other parameters:** *Self-avoiding* in the *crossover point* and *Elitism* not activated.

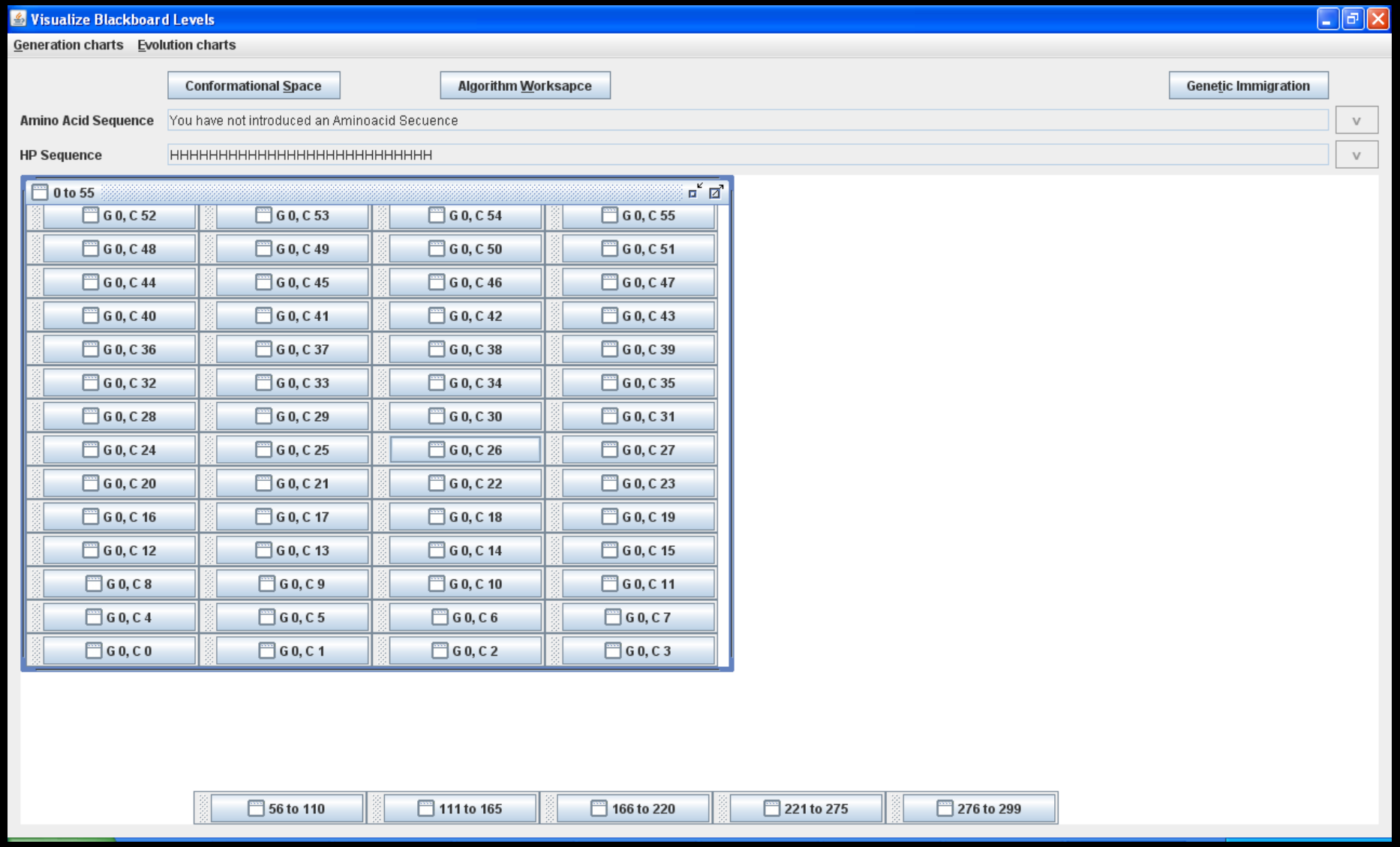

Computational Framework for Modeling and Simulation of Biological Problems
Project
Run
Tools
About
New Project
Open Project
Save Alt-S
Save As... Alt+Mayúsculas-S
Close Alt-C
Exit Alt-X
Evolution
Framework for Protein Folding Problems
Intrtoduces the Aminoacid or HP secuence
What kind of sequence do you want to enter?
HP Sequence
Amino Acid Sequence
HHHHHHHHHHHHHHHHHHHHHHHHHH
OK
From File


Computational Framework fo deling and Simulation of Biological Problems
Project
Run
Tools
About
Run Application Alt-R
View Results Alt-V
Evolution
Framework for Protein Folding Problems
Run Application
Select the Model and Algorithm to Run:
HP Sequence:
HP Generator
Lattice:
Lattice
Lattice
2D Square
2D Triangle
3D Square
3D Diamond
Run Genetic Algorithm
Partial Progress
0%
Total Progress
0%
Conformational Space
Choose the number of conformations to generate:
300
Aceptar
Cancelar

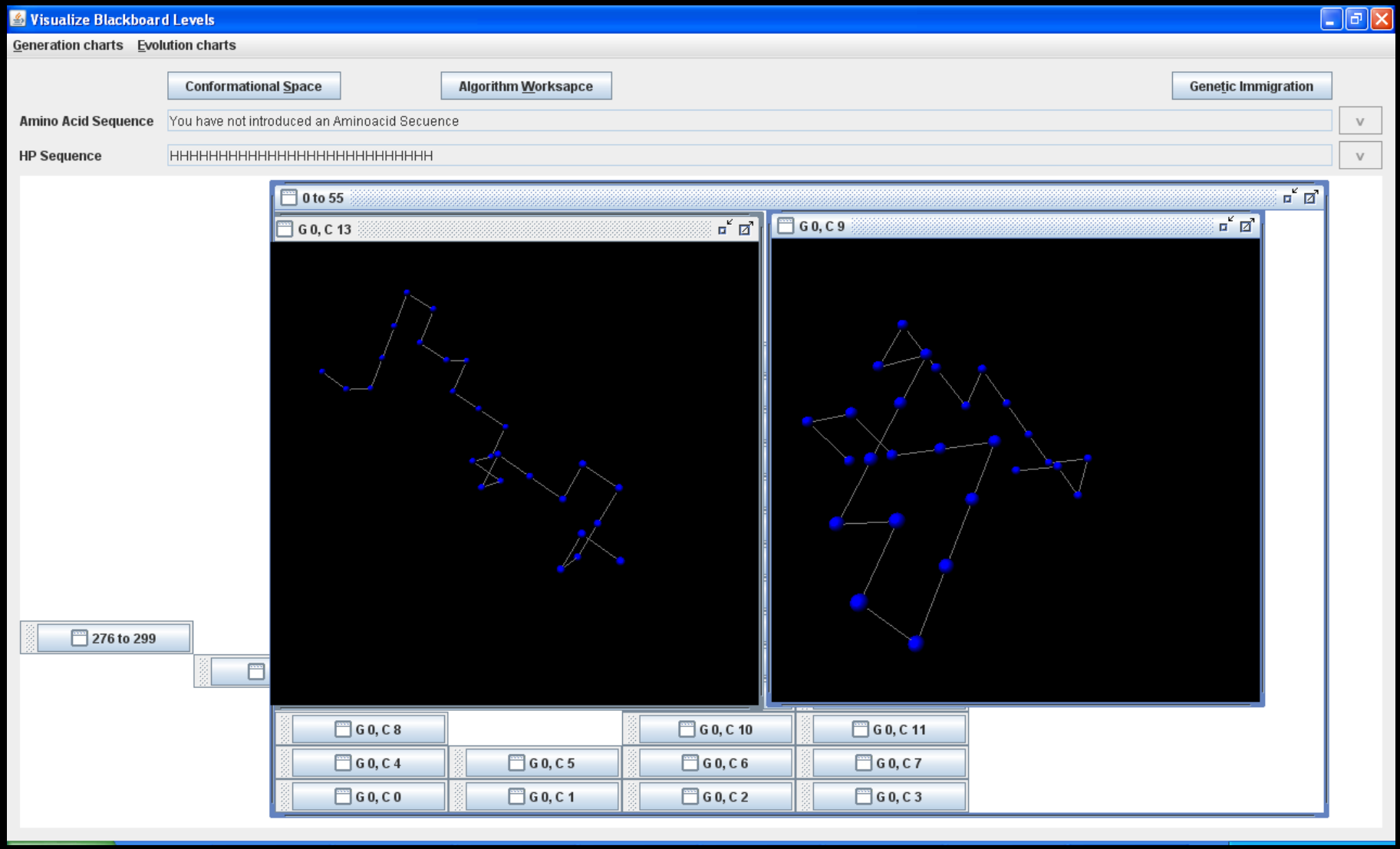

Computational Framework for Modeling and Simulation of Biological Problems
Project
Run
Tools
About
Run Application Alt-R
View Results Alt-V
Evolution
Framework for Protein Folding Problems

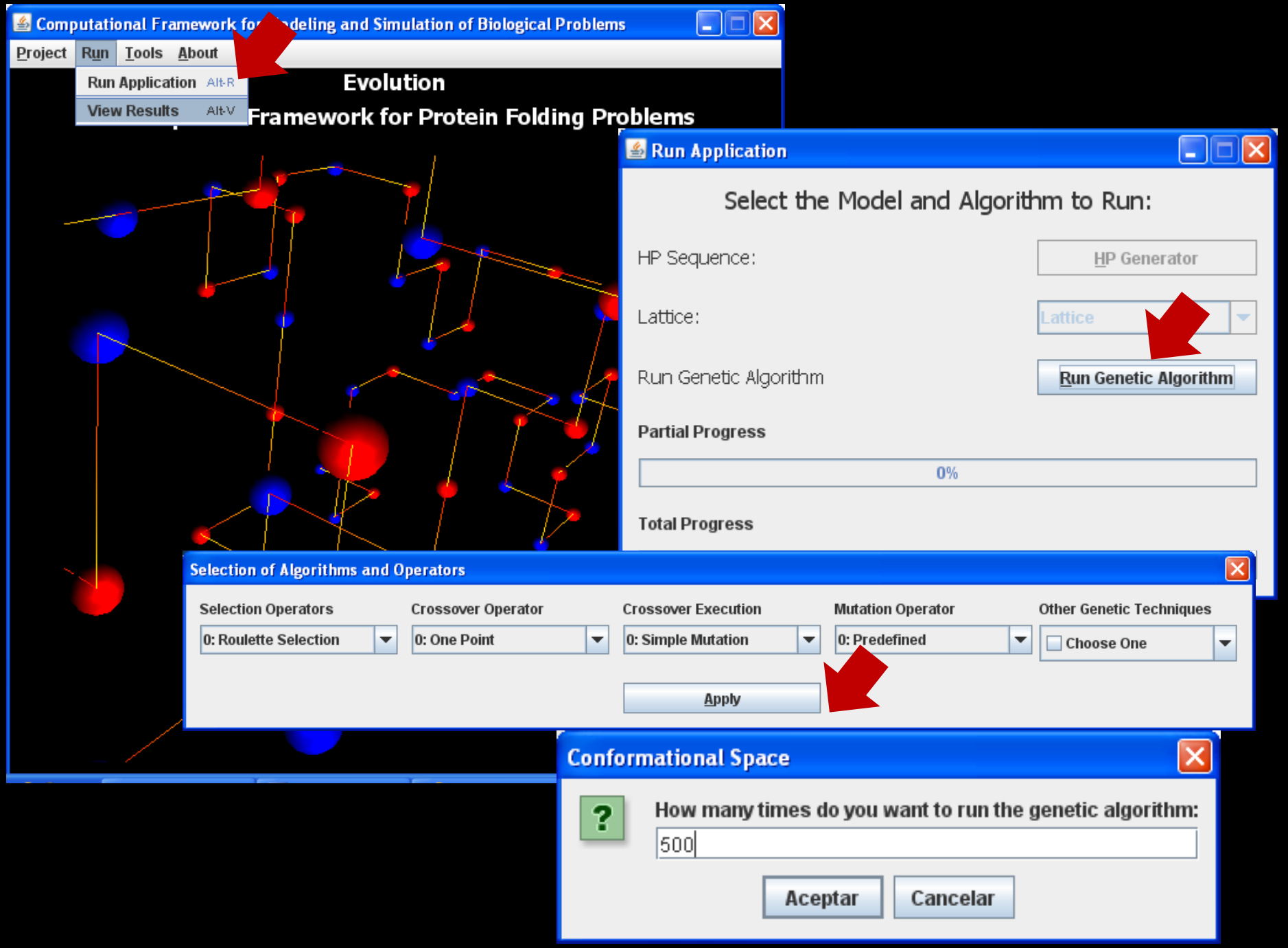

Visualize Blackboard Levels
Generation charts
Evolution charts
Conformational Space
Algorithm Worksapce
Genetic Immigration
Amino Acid Sequence
You have not introduced an Aminoacid Secuence
HP Sequence
HHHHHHHHHHHHHHHHHHHHHHHHHH
PLATAFOR...
DISENO BI...
INFORMES ...
Informe de ...
Evolution S...
ProteinFoldi...
Computatio...

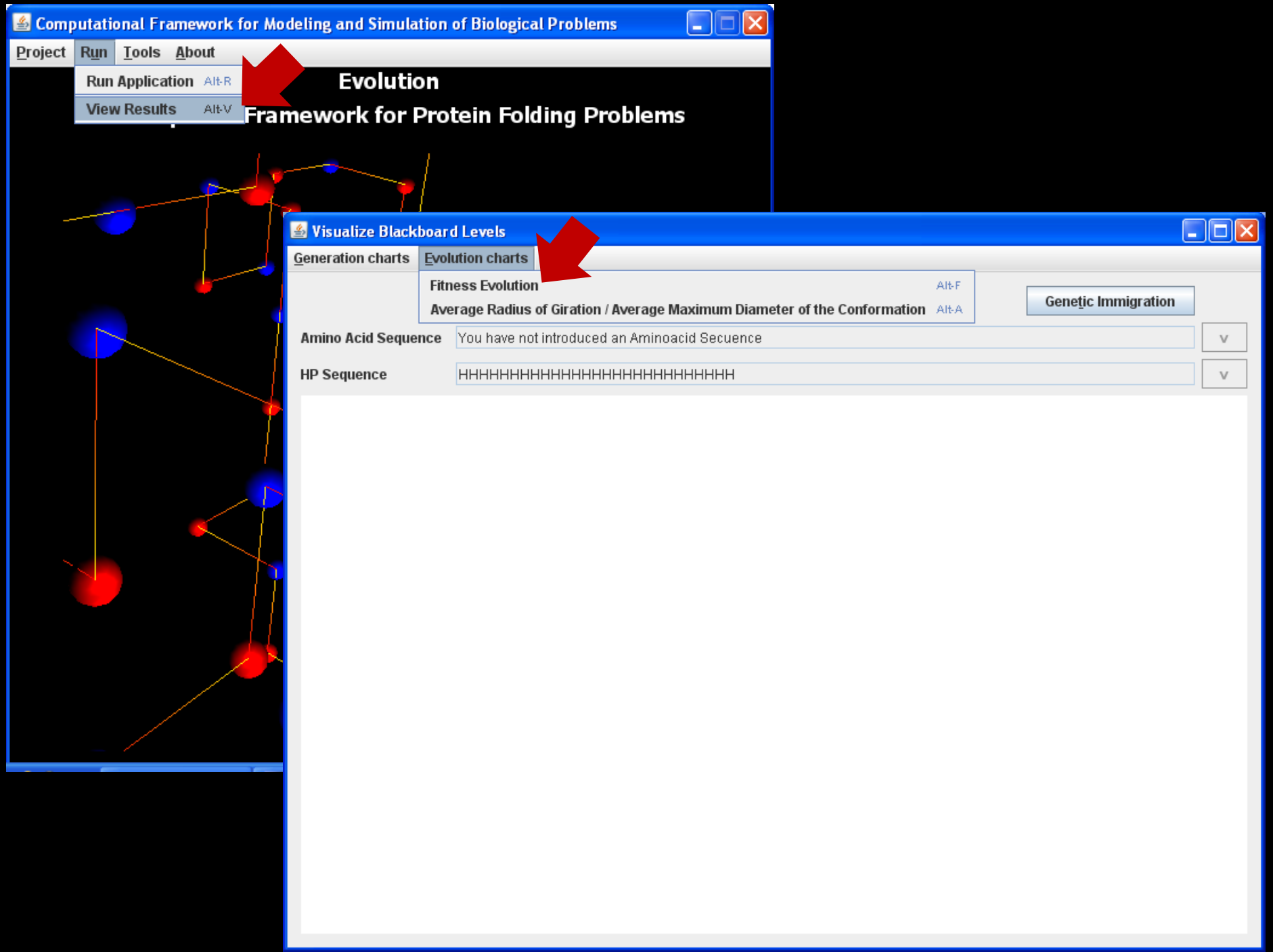
The initial generation

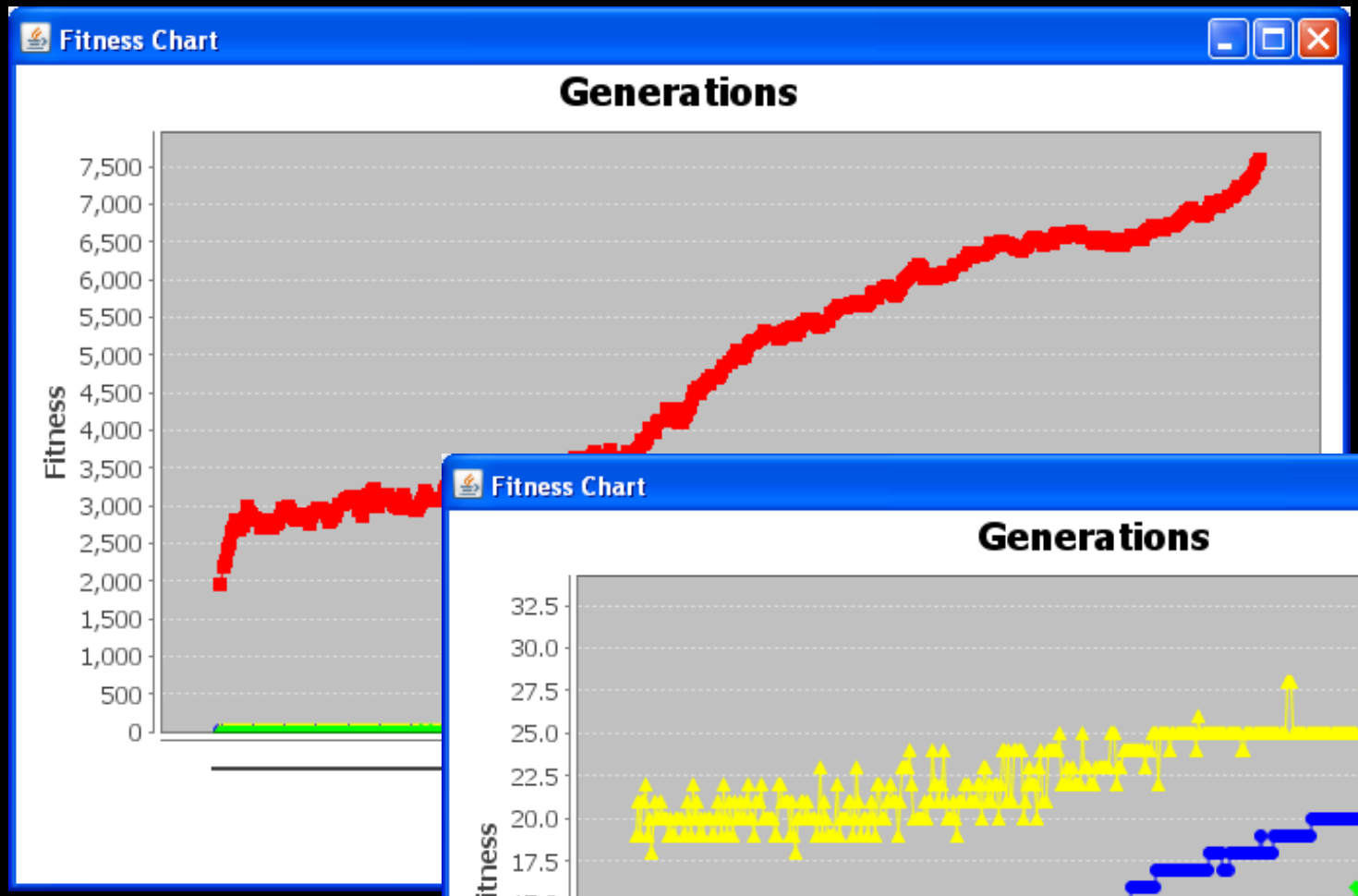


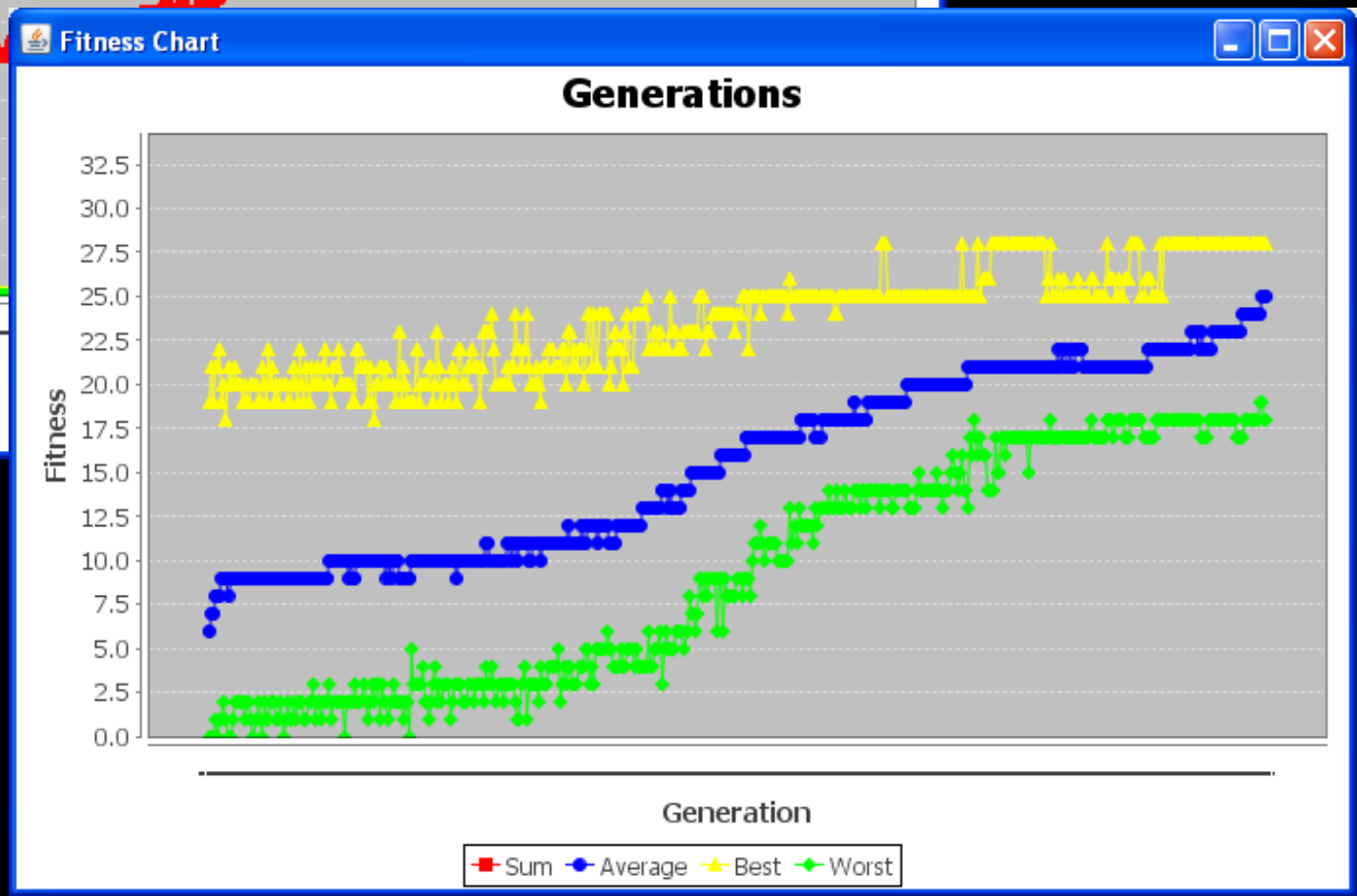
Two members from the initial generation

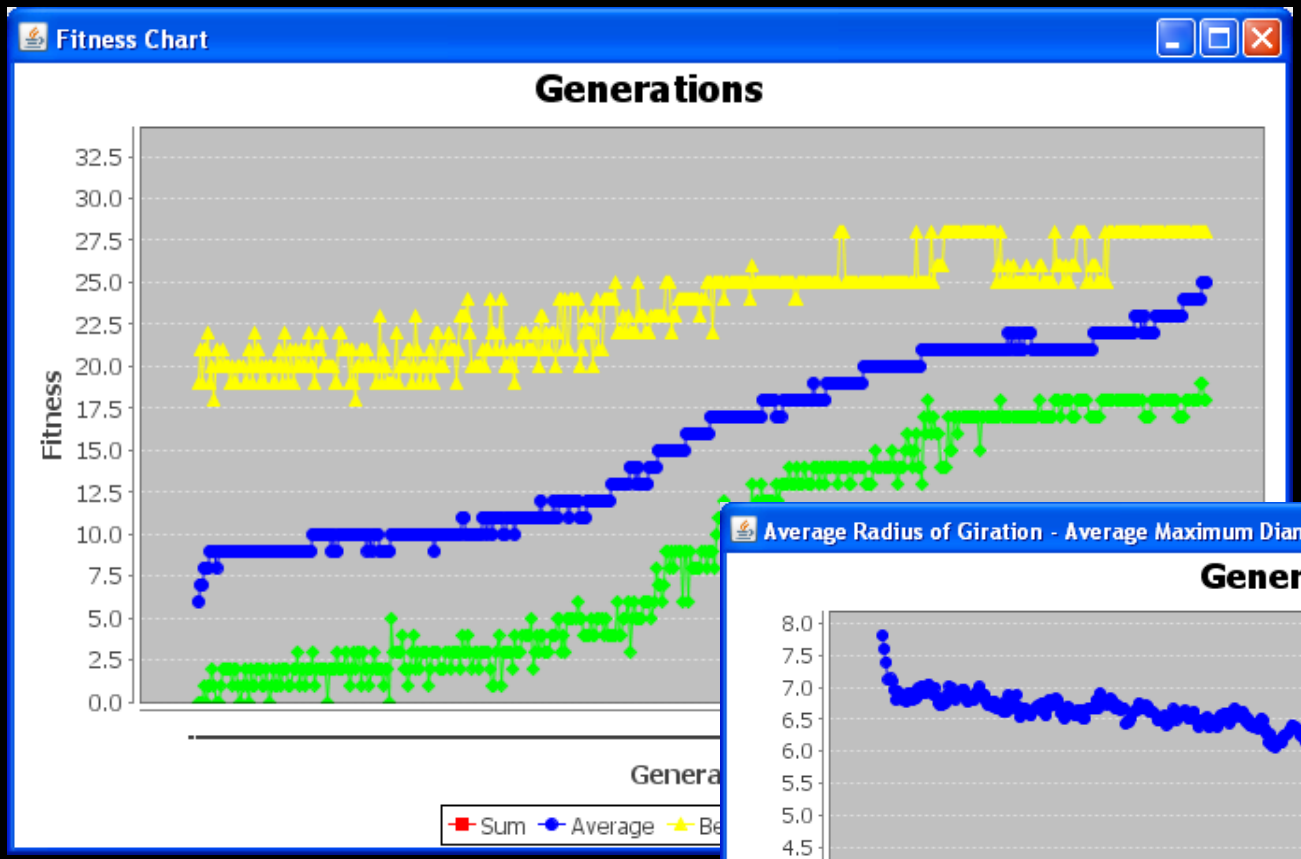


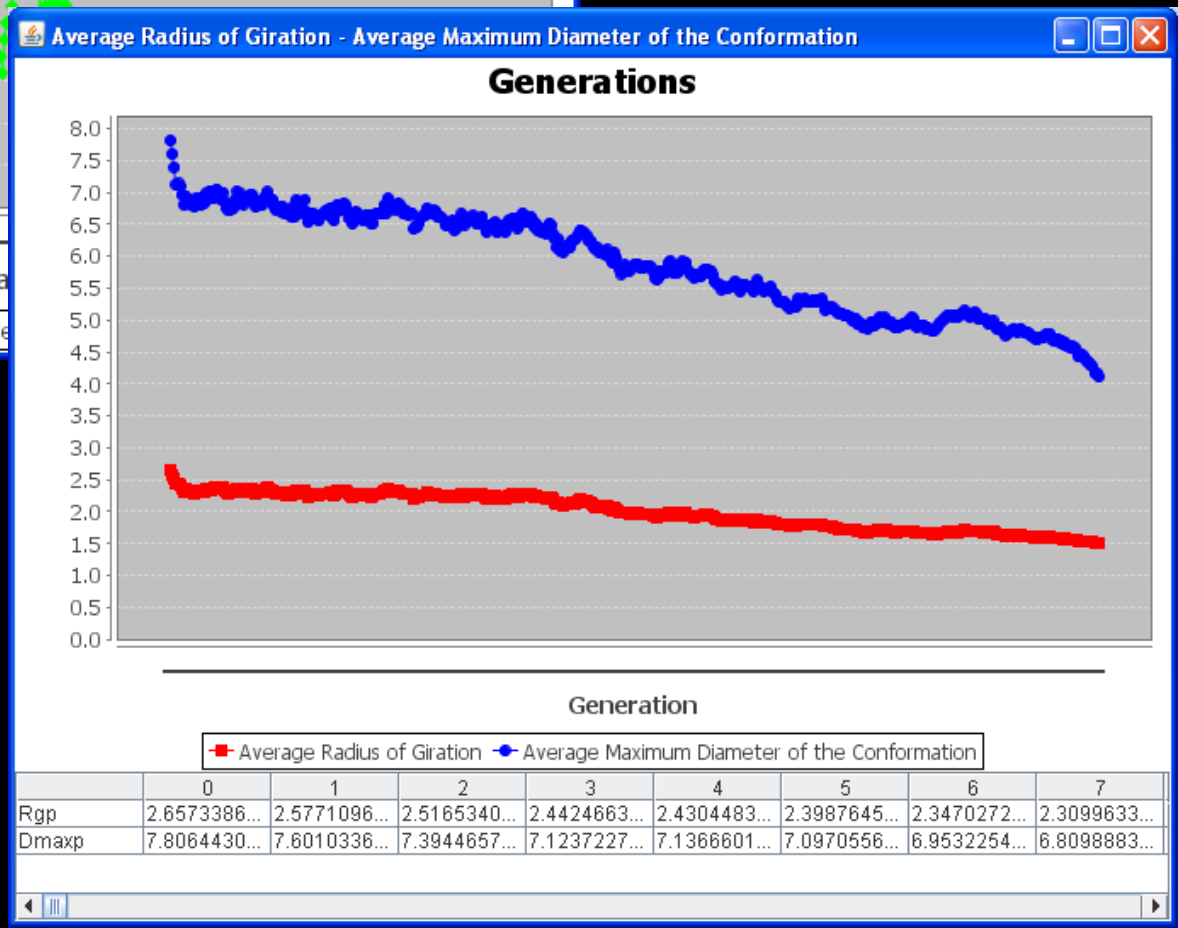

Computational Framework for Modeling and Simulation of Biological Problems
Project
Run
Tools
About
Run Application Alt-R
View Results Alt-V
Evolution
Framework for Protein Folding Problems
Run Application
Select the Model and Algorithm to Run:
HP Sequence:
HP Generator
Lattice:
Lattice
Run Genetic Algorithm
Run Genetic Algorithm
Partial Progress
0%
Total Progress
Selection of Algorithms and Operators
Selection Operators
0: Roulette Selection
Crossover Operator
0: One Point
Crossover Execution
0: Simple Mutation
Mutation Operator
0: Predefined
Other Genetic Techniques
Choose One
Apply
Conformational Space
How many times do you want to run the genetic algorithm:
500
Aceptar
Cancelar


Computational Framework for Modeling and Simulation of Biological Problems
Project
Run
Tools
About
Run Application Alt-R
View Results Alt-V
Evolution
Framework for Protein Folding Problems
Visualize Blackboard Levels
Generation charts
Evolution charts
Fitness Evolution Alt-F
Average Radius of Giration / Average Maximum Diameter of the Conformation Alt-A
Genetic Immigration
Amino Acid Sequence
You have not introduced an Aminoacid Secuence
HP Sequence
HHHHHHHHHHHHHHHHHHHHHHHHHH

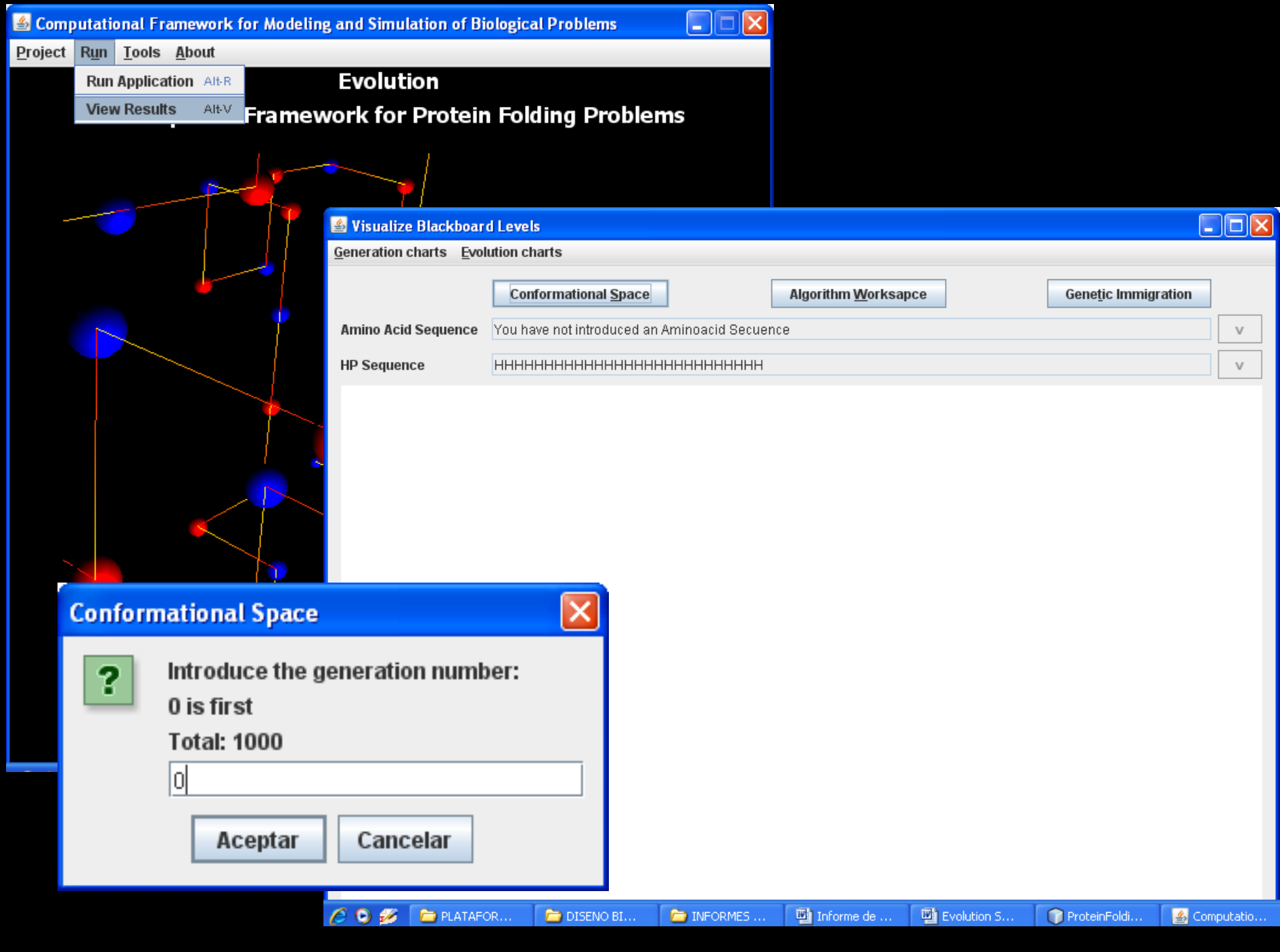

Fitness Chart
Generations
Fitness
7,500
7,000
6,500
6,000
5,500
5,000
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0

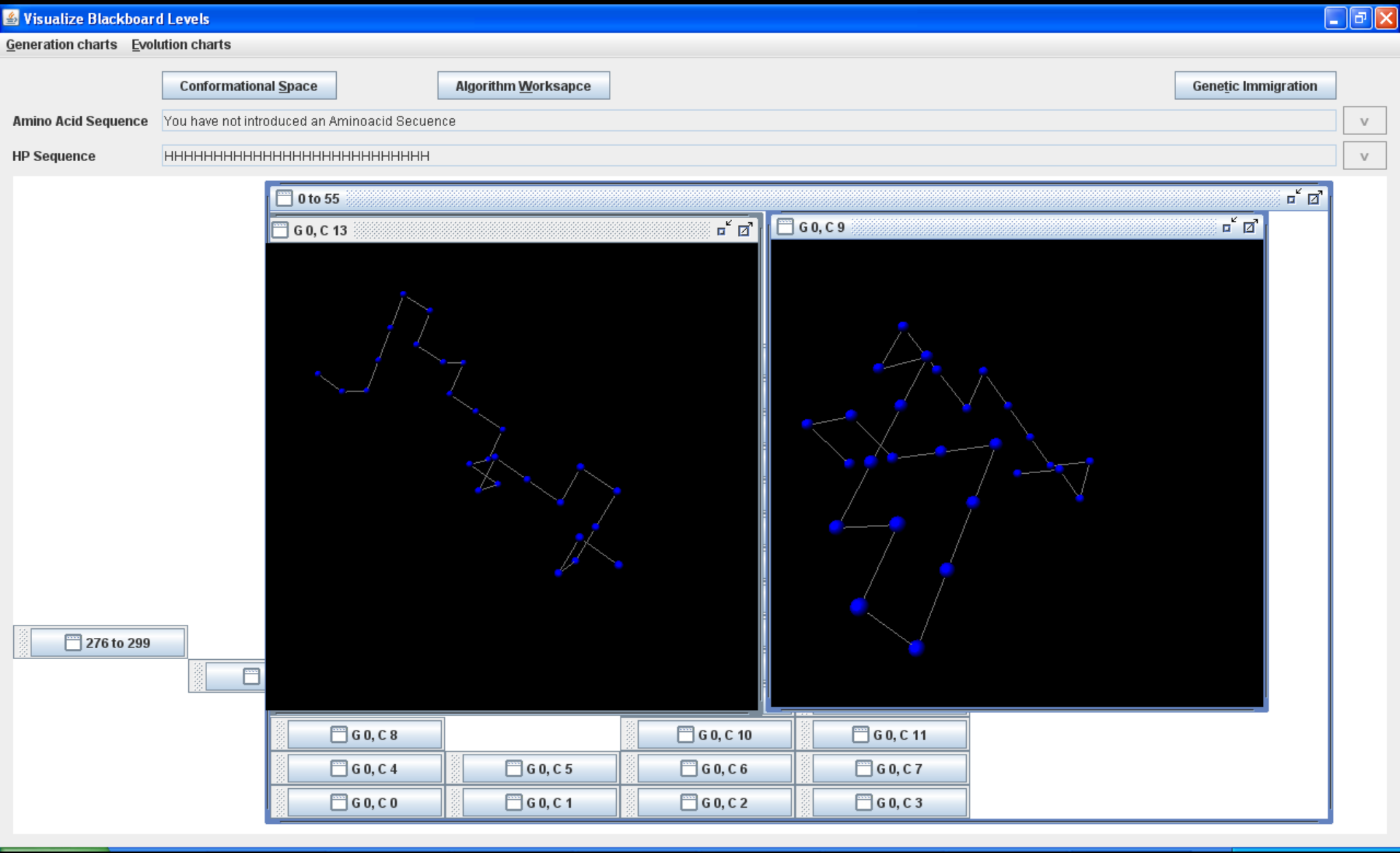

Fitness Chart
Generations
Fitness
32.5
30.0
27.5
25.0
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Generation
Sum
Average
Best
Worst



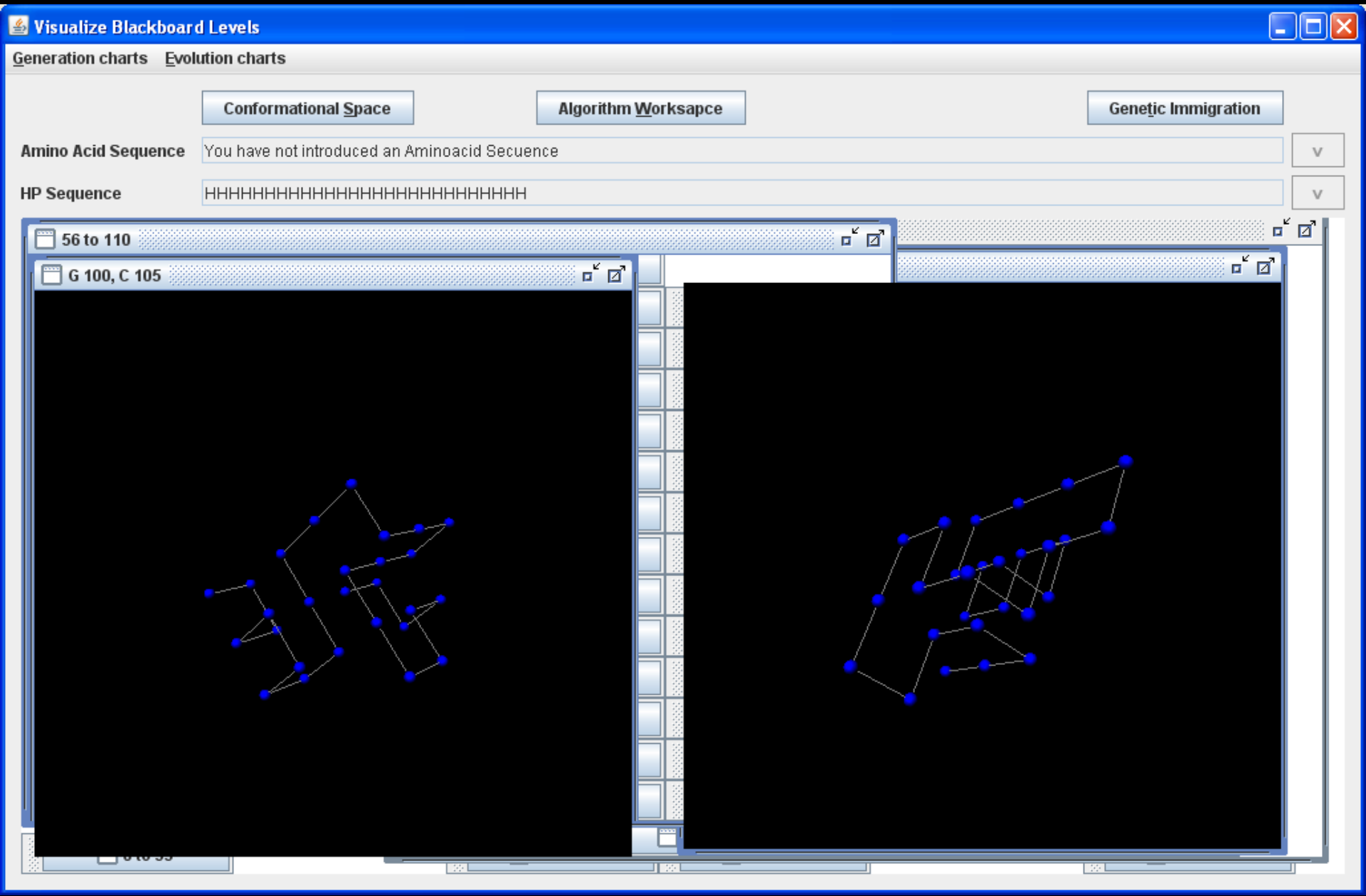


	0	1	2	3	4	5	6	7
Rgp	2.6573386...	2.5771096...	2.5165340...	2.4424663...	2.4304483...	2.3987645...	2.3470272...	2.3099633...
Dmaxp	7.8064430...	7.6010336...	7.3944657...	7.1237227...	7.1366601...	7.0970556...	6.9532254...	6.8098883...

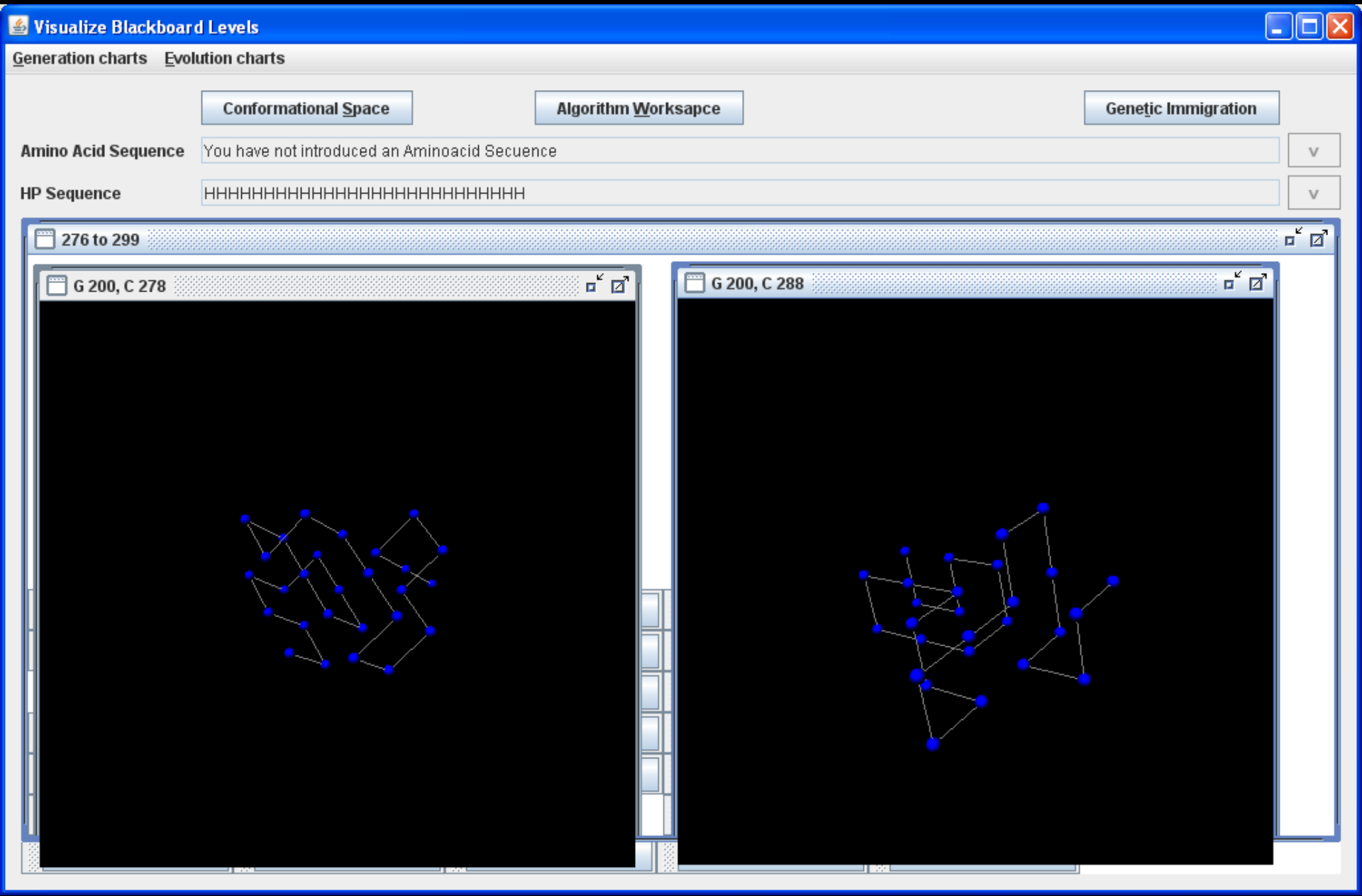

Computational Framework for Modeling and Simulation of Biological Problems
Project
Run
Tools
About
Run Application Alt-R
View Results Alt-V
Evolution
Framework for Protein Folding Problems
Visualize Blackboard Levels
Generation charts
Evolution charts
Conformational Space
Algorithm Worksapce
Genetic Immigration
Amino Acid Sequence
You have not introduced an Aminoacid Secuence
HP Sequence
HHHHHHHHHHHHHHHHHHHHHHHHHH
Conformational Space
Introduce the generation number:
0 is first
Total: 1000
0
Aceptar
Cancelar
PLATAFOR...
DISENO BI...
INFORMES ...
Informe de ...
Evolution S...
ProteinFoldi...
Computatio...



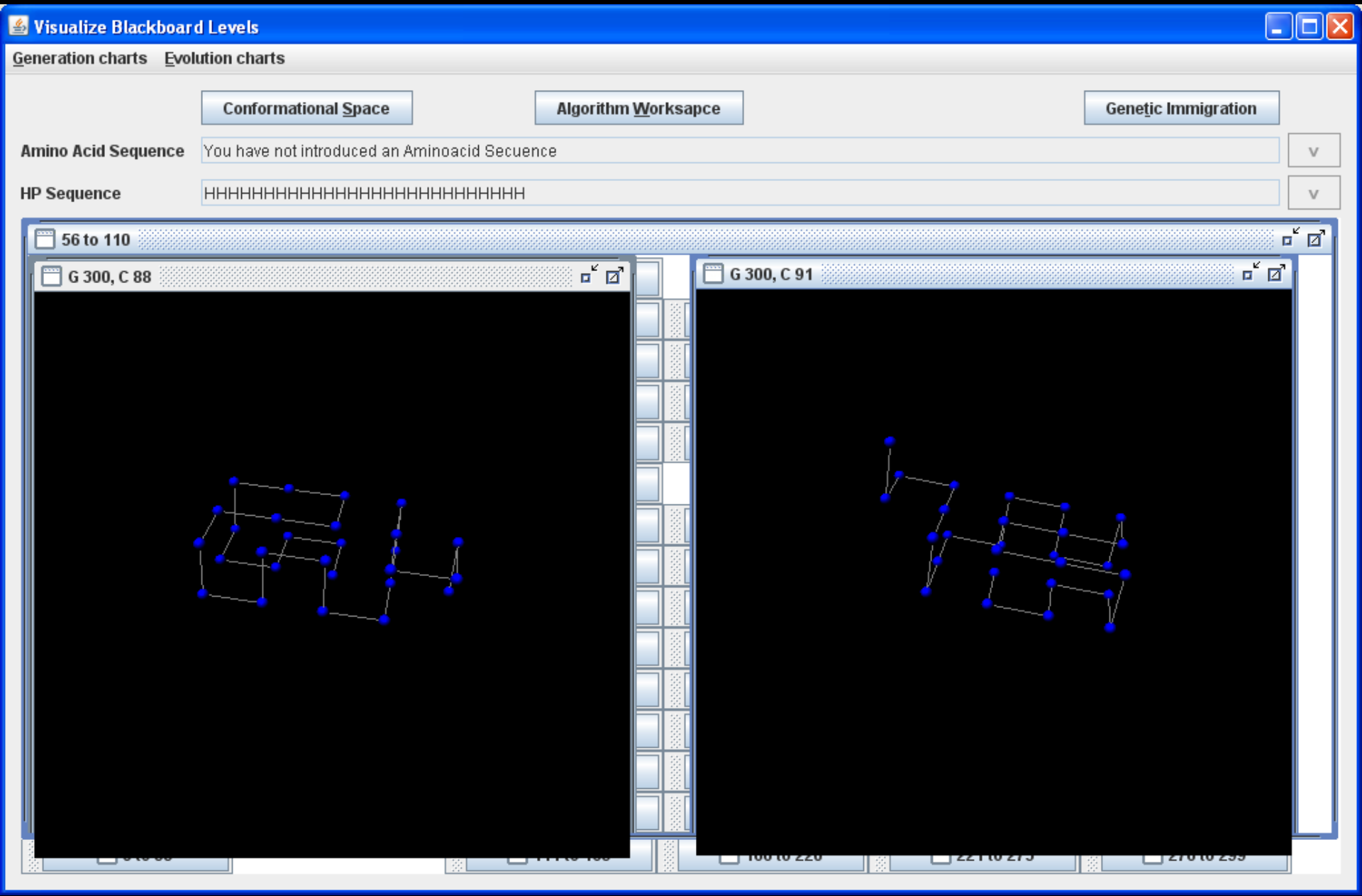
Two members from the initial generation



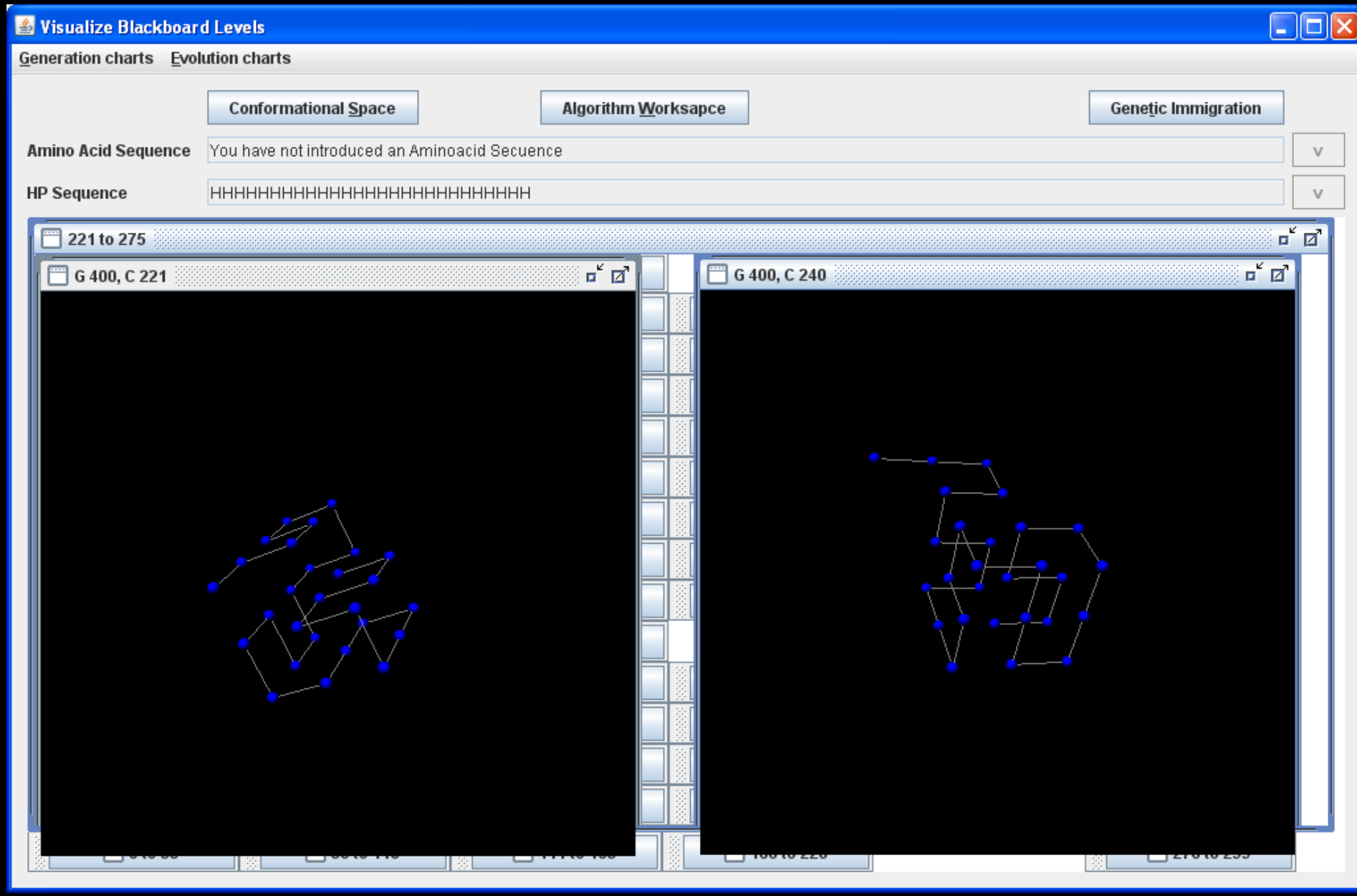
Two members from generation 100



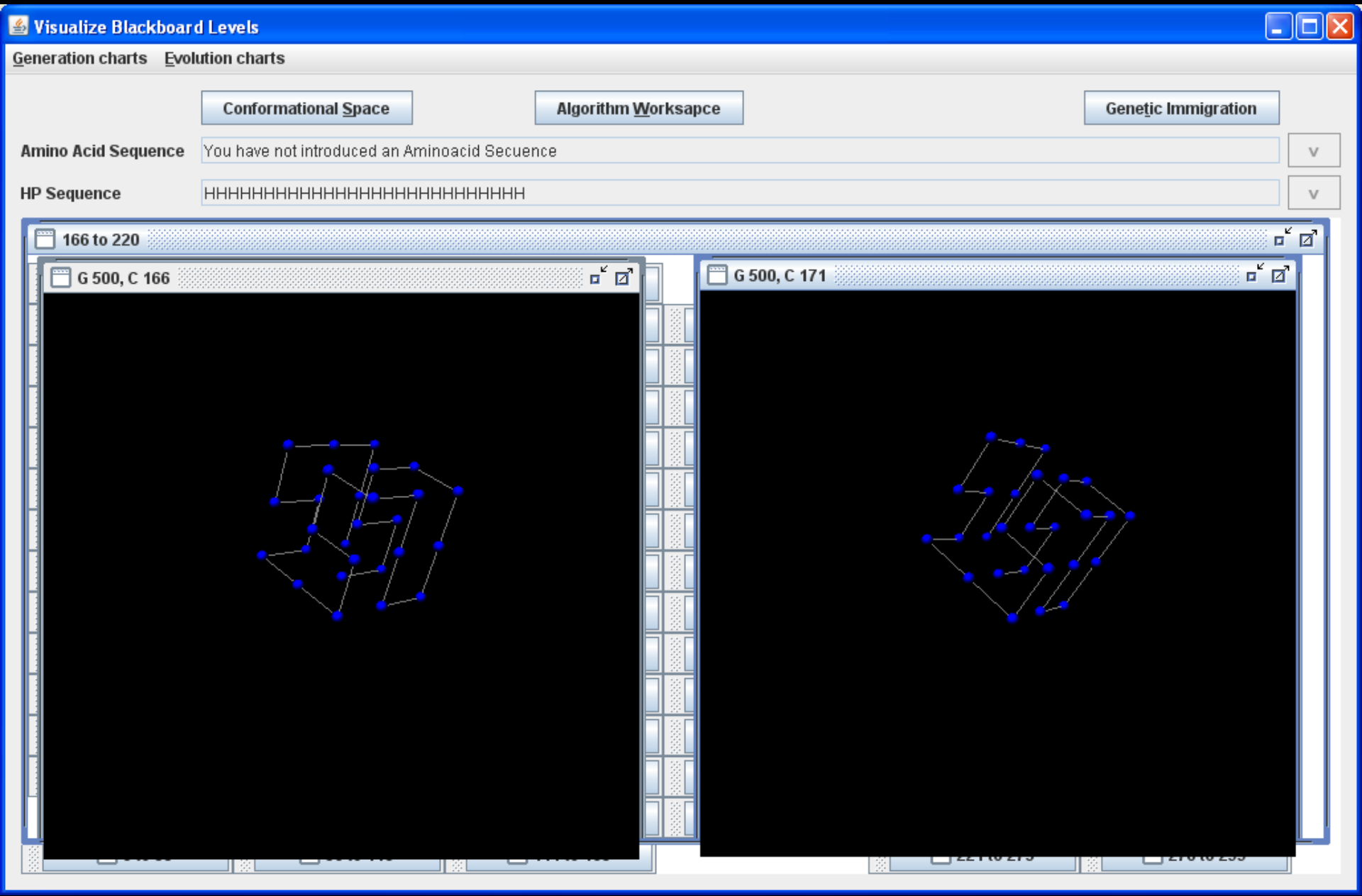
Two members from generation 200



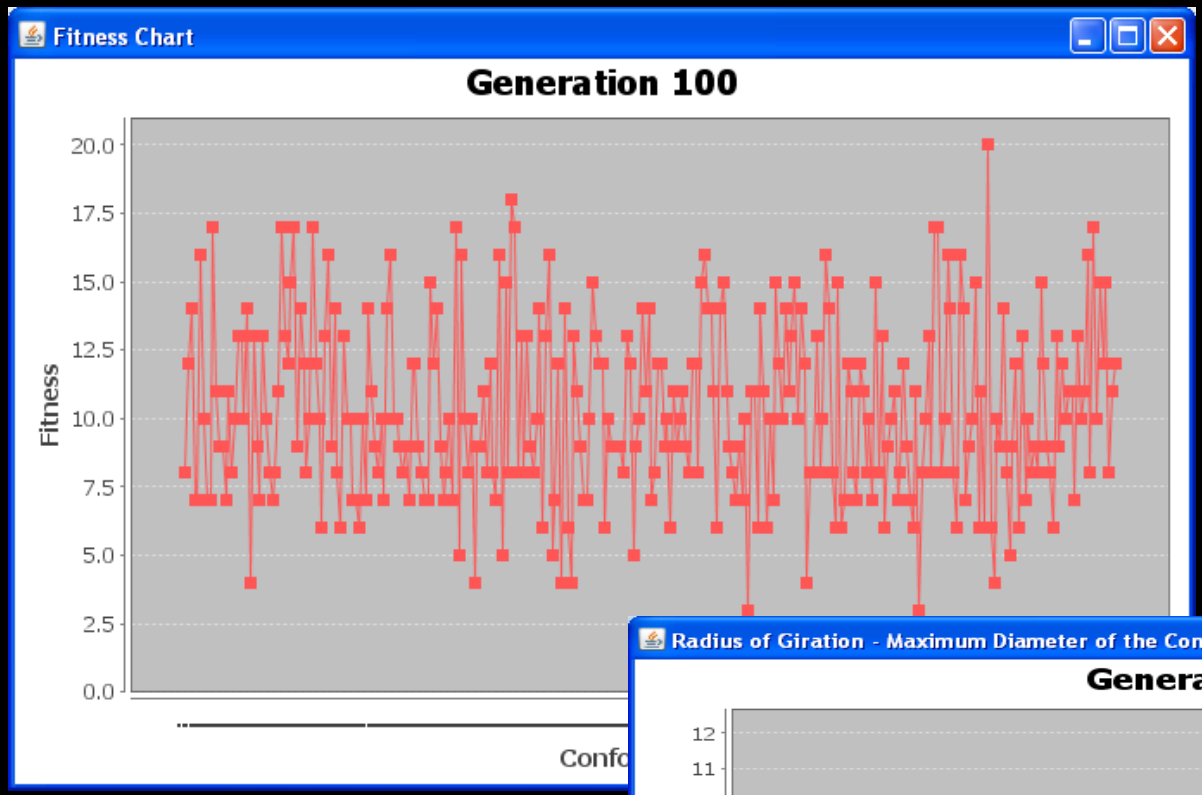
Two members from generation 300

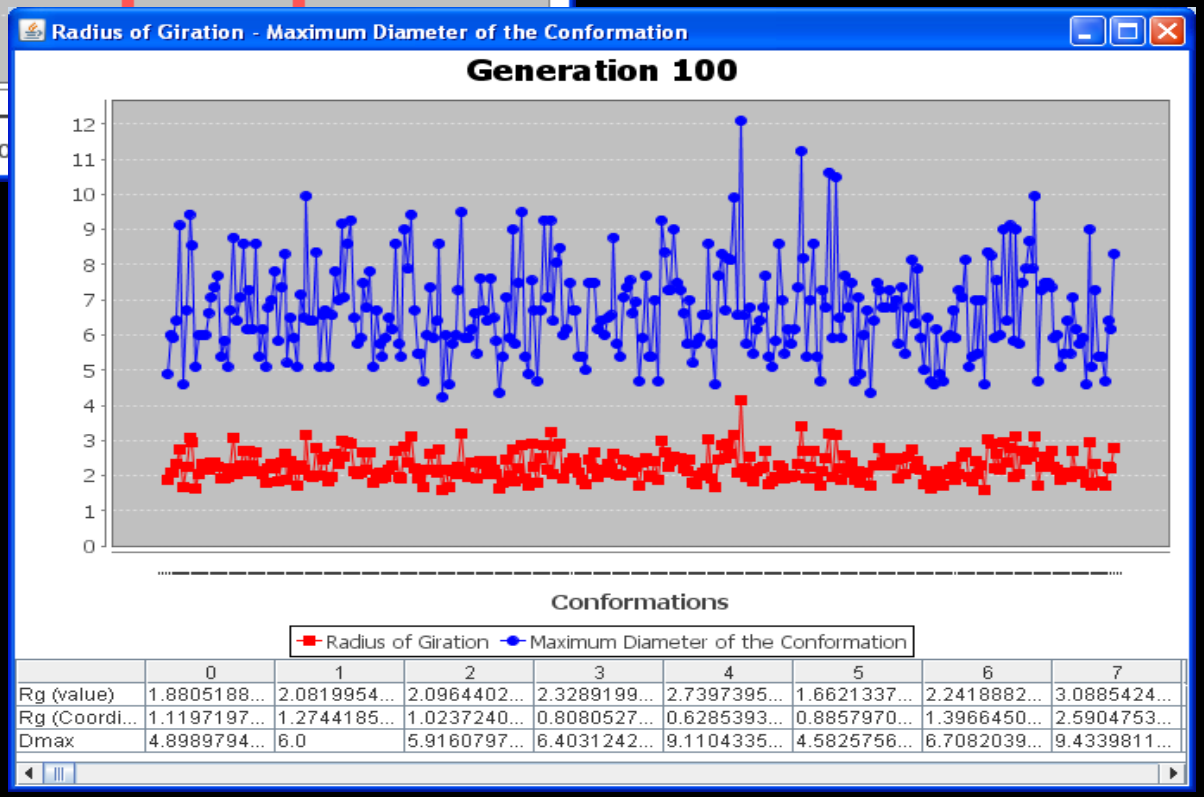


Two members from generation 400

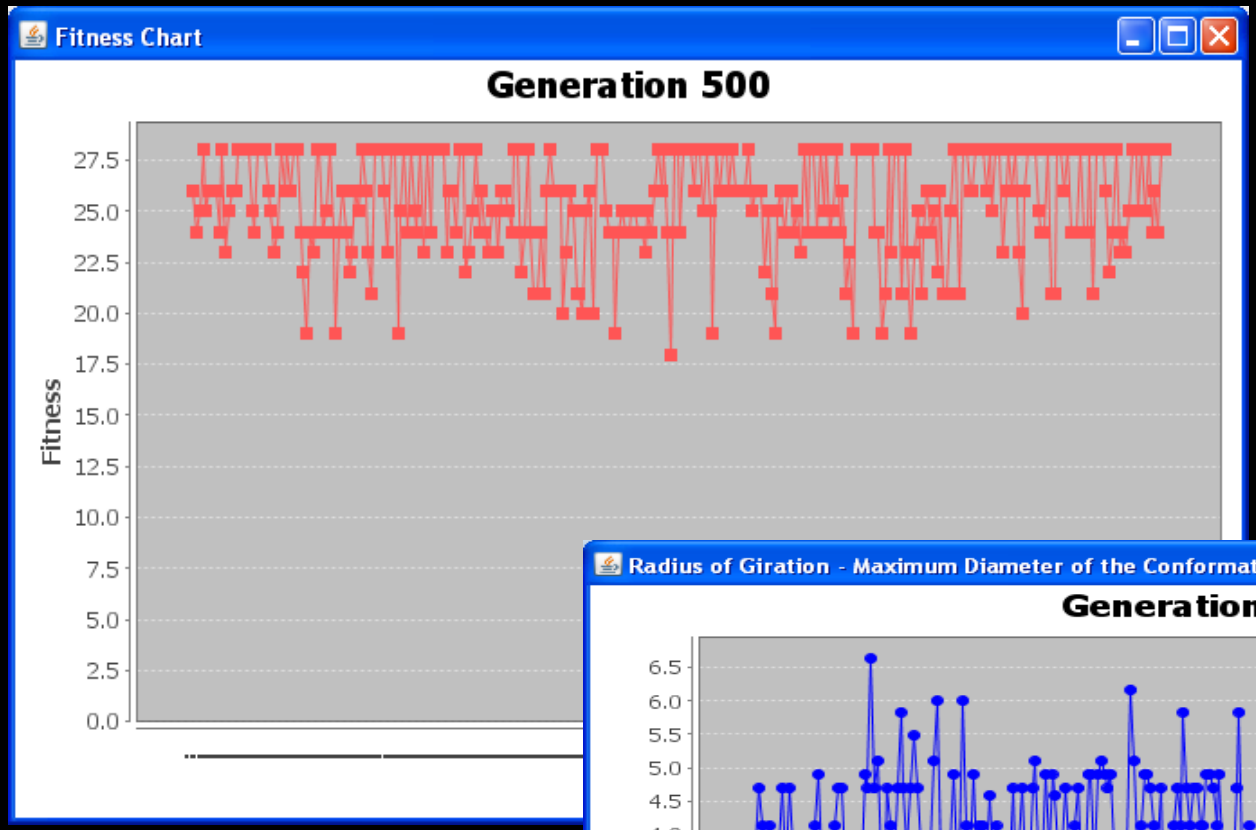


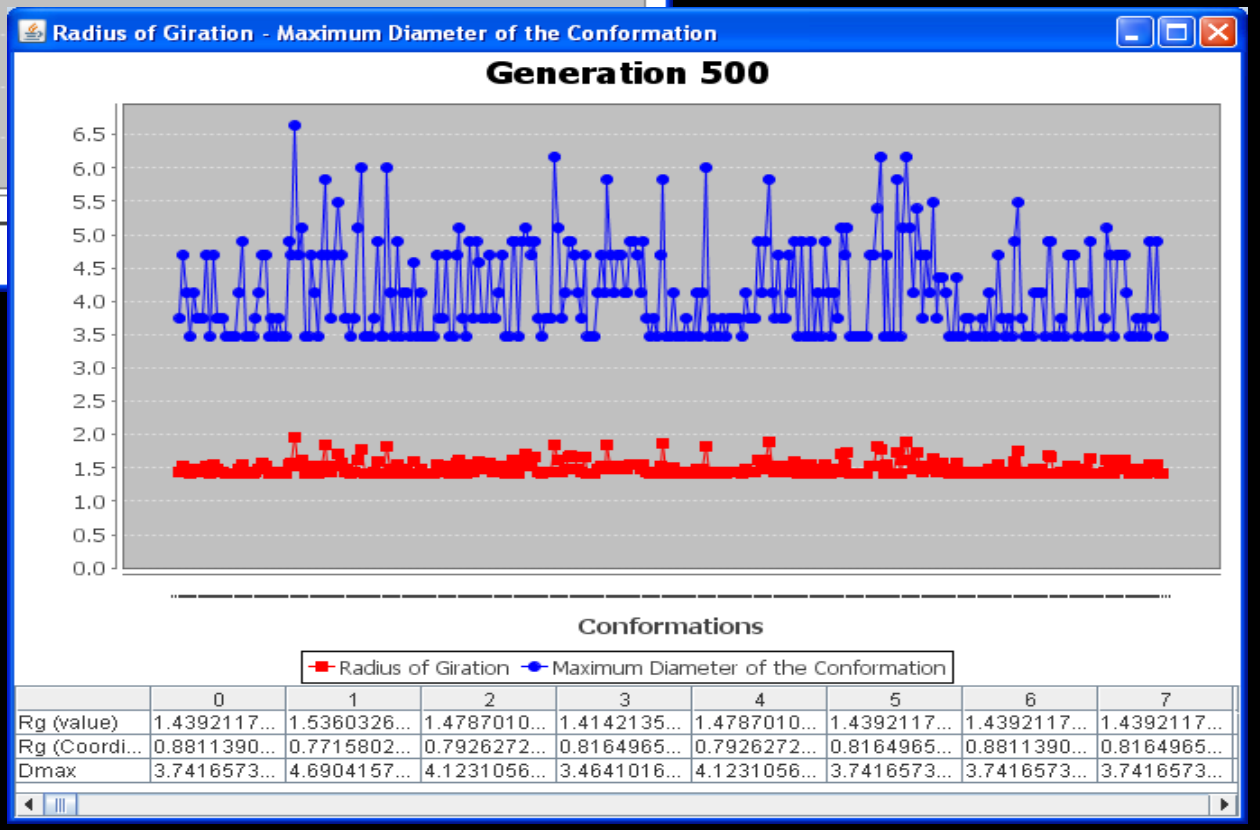
Two members from generation 500





	0	1	2	3	4	5	6	7
Rg (value)	1.8805188...	2.0819954...	2.0964402...	2.3289199...	2.7397395...	1.6621337...	2.2418882...	3.0885424...
Rg (Coordi...	1.1197197...	1.2744185...	1.0237240...	0.8080527...	0.6285393...	0.8857970...	1.3966450...	2.5904753...
Dmax	4.8989794...	6.0	5.9160797...	6.4031242...	9.1104335...	4.5825756...	6.7082039...	9.4339811...

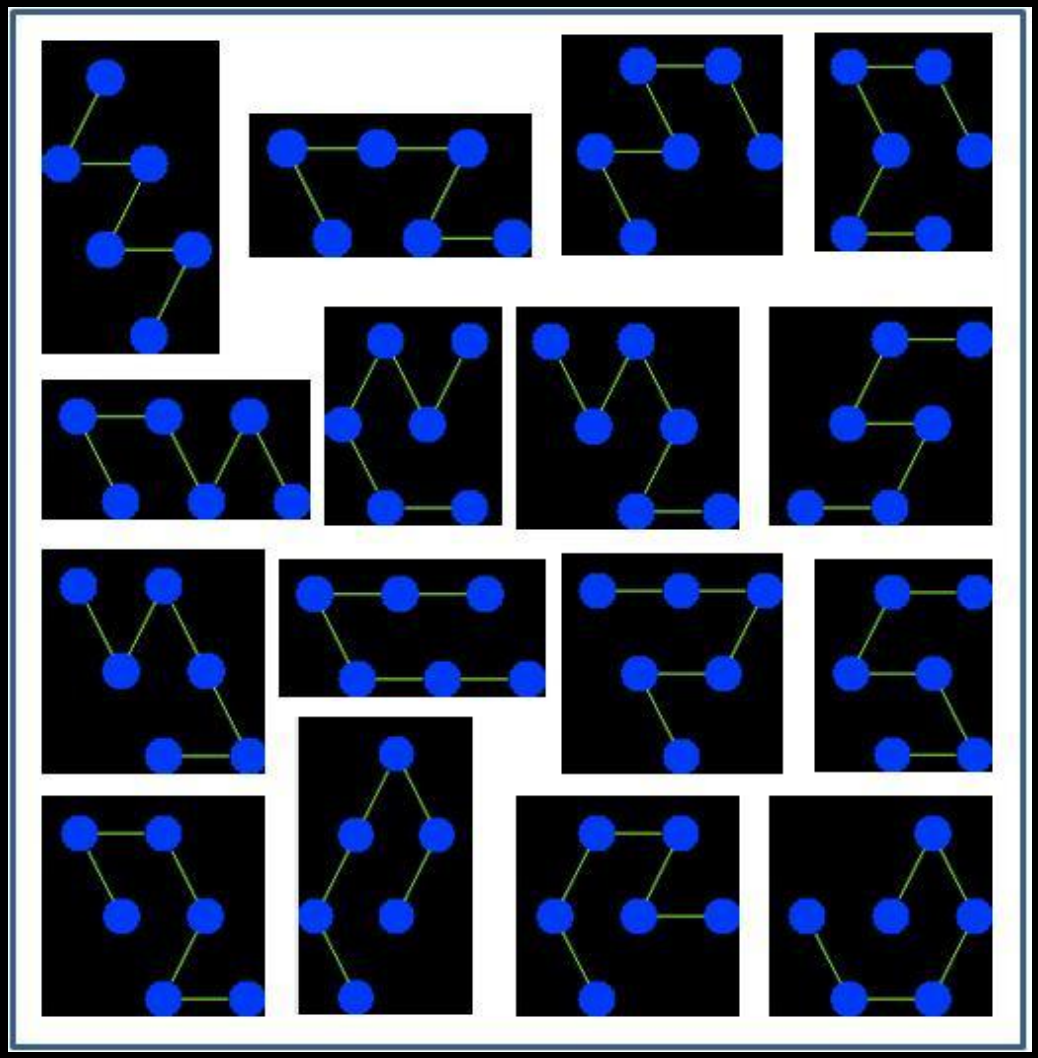




	0	1	2	3	4	5	6	7
Rg (value)	1.4392117...	1.5360326...	1.4787010...	1.4142135...	1.4787010...	1.4392117...	1.4392117...	1.4392117...
Rg (Coordi...	0.8811390...	0.7715802...	0.7926272...	0.8164965...	0.7926272...	0.8164965...	0.8811390...	0.8164965...
Dmax	3.7416573...	4.6904157...	4.1231056...	3.4641016...	4.1231056...	3.7416573...	3.7416573...	3.7416573...

Search of local determinants of structure

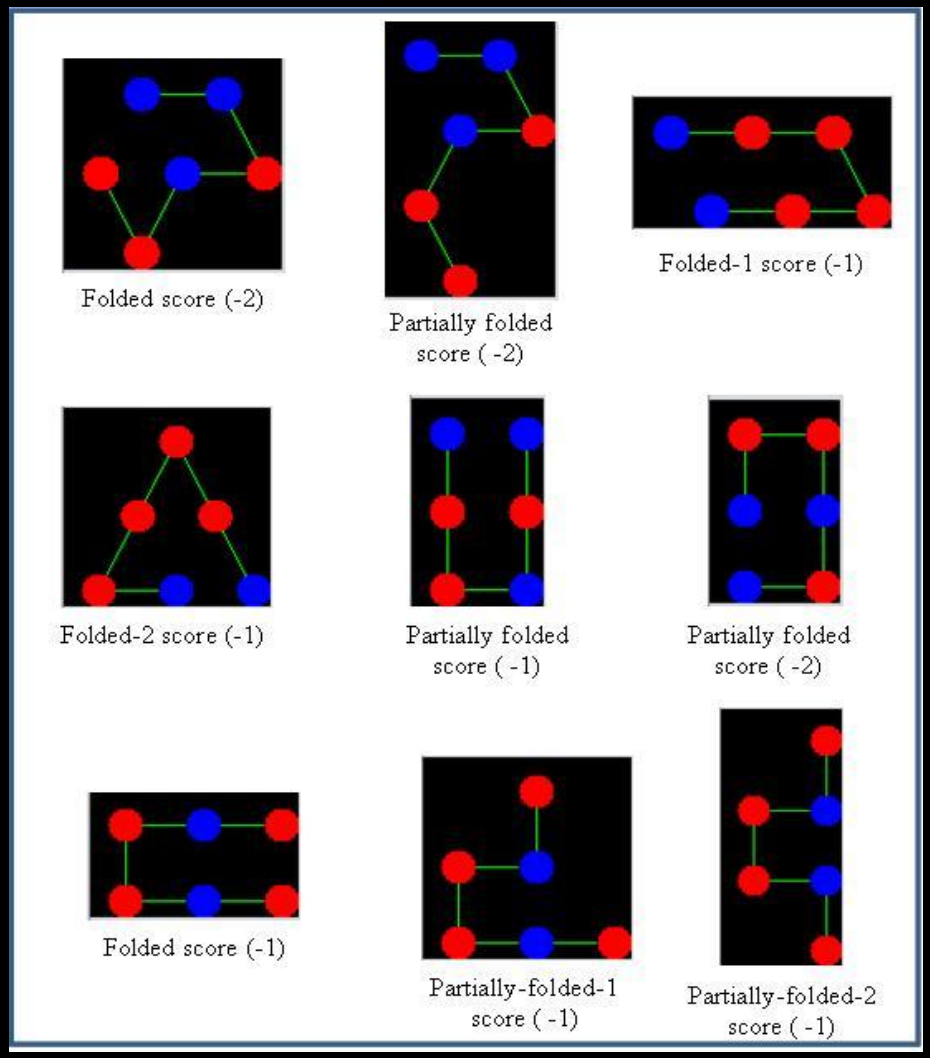
SELECTED SEQUENCES OF HEXAMERIC HP SYSTEMS	
HHHHHH	PPHPHH
PHHHHH	PPHHPH
HPHHHH	PHPPHH
HHPHHH	HPPPHH
PPHHHH	HPHPPH
PHPHHH	PHPHPH
PHHPHH	HPPHPH
PHHHPH	HHPPPH
PHHHHP	HPPPPH
HPPHHH	HPPHPP
HPHPHH	PHPPHP
HPHHPH	PHHPPH
HHPPHH	PPPPPP



Hexamer folding in 2D triangular lattice. For the HHHHHH several degenerated structures, with the maximum score of 4 can be detected with the use of the platform. A minus 1 is added to the score for each H bead with another in a close neighbor point in the lattice with no bond between them. Meanwhile for the other sequences varying degrees of degeneracy and maximum number of contacts can be tracked.

Search of local determinants of structure

SELECTED SEQUENCES OF HEXAMERIC HP SYSTEMS	
HHHHHH	PPHPHH
PHHHHH	PPHHPH
HPHHHH	PHPPHH
HHPHHH	HPPPHH
PPHHHH	HPHPPH
PHPHHH	PHPHPH
PHHPHH	HPPHPH
PHHHPH	HHPPPH
PHHHHP	HPPPPH
HPPHHH	HPPHPP
HPHPHH	PHPPHP
HPHHPH	PHHPPH
HHPPHH	PPPPPP



Hexamer folding in 2D square and triangular lattice. Note that with the current fitting function the same optimal score can be obtained with close packed (globular) chains and with low density (relatively extended) ones, even for short sequences as the conformers shown in this figure. This suggest the inclusion of packing indexes in the fitting function.

New molecular indexes Radius of gyration and Maximal diameter

$$R_g = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(r_k - r_{mean}\right)^2}$$

$$\forall r_i, r_j \quad i, j = 1,2,\cdots,N, \quad where$$

$$r_i = \left(x_i, y_i, z_i\right), \quad r_j = \left(x_j, y_j, z_j\right)$$

$$calculate \quad d_{ij} = d\left(r_i, r_j\right)$$

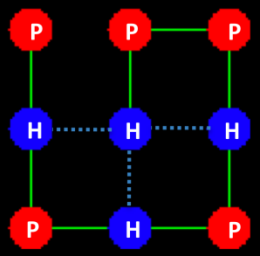
$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2}$$

$$D_{Max} = \max\left(d_{ij}\right)$$

$$F = f(e) + g\left(R_g\right) + h\left(D_{Max}\right)$$

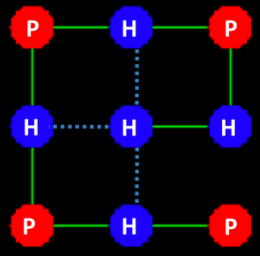
Algorithm performance

Sequence 1
HPPHPHPHP



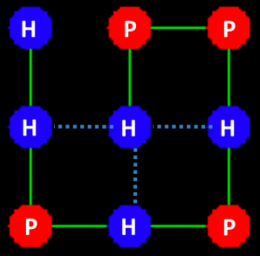
Score -3
4H/5P

Sequence 2
HHPHPHPHP



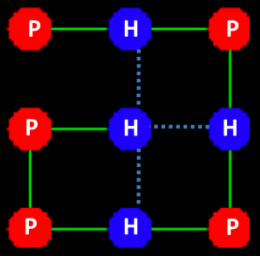
Score -3
5H/4P

Sequence 3
HPPHPHPHH



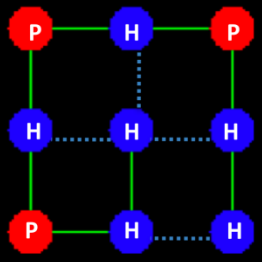
Score -3
5H/4P

Sequence 4
HPPHPHPHP



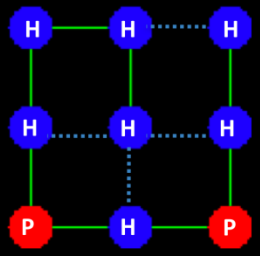
Score -3
4H/5P

Sequence 5
HHPHPHPHH



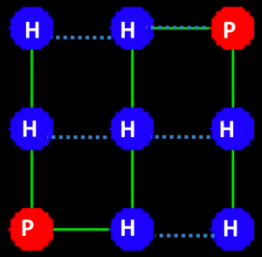
Score -4
6H/3P

Sequence 6
HHHHPHPHH



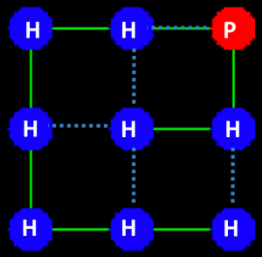
Score -4
7H/2P

Sequence 7
HHPHHHPHH



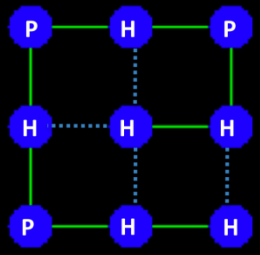
Score -4
7H/2P

Sequence 8
HHHHHHPHH



Score -4
8H/1P

Sequence 9
HHHHHHHHH



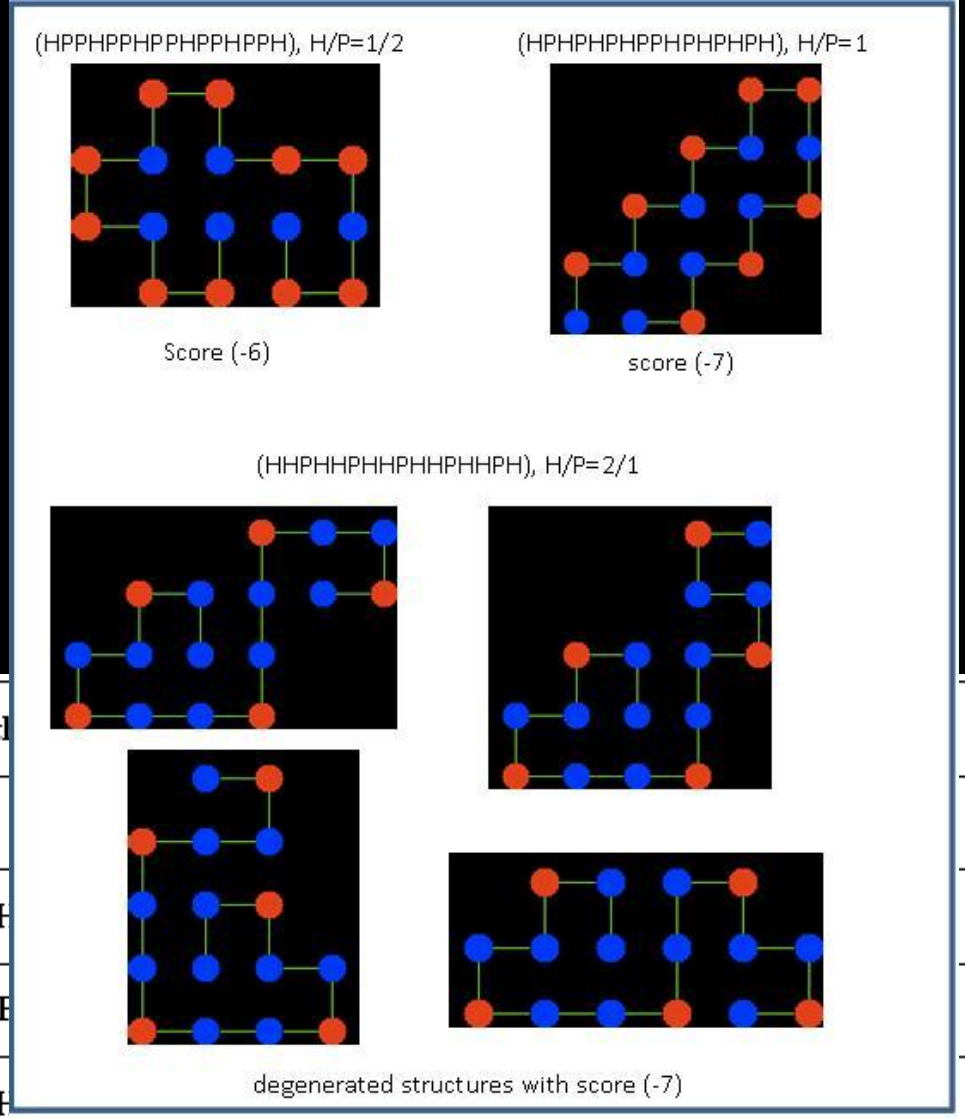
Score -4
9H/0P

Folded conformations of varying sequences differing not only in their H/P ratio but also in the positioning of the residues. The sequences were constructed trying to yield the highest scores possible upon folding into a 2D square, in order to test the performance of the platform.

*Looking for **optimal generation size** and **minimum number of generations** in a 2D-squared lattice*

Some of the HP sequences assayed			
L	HP ≈ 1/2	HP = 1	HP ≈ 2/1
10	HPPHPPHPPH	HPHPHPPHPH	HHPHHPHPHH
12	HPPPHPPHPPPH	HPHPHPPHPHPH	HPHHPHHPHHPH
14	HPPHPPHPPHPPPH	HPHPHPHPPHPHPH	HHPHHPHHPHHPHH
16	HPPHPPHPPHPPHPPH	HPHPHPHPPHPHPHPH	HHPHHPHHPHHPHHPH
18	HPPHPPPHPPHPPHPPPH	HPHPHPHPHPPHPHPHPH	HPHHPHHPHHPHHPHHPH
20	HPPHPPHPPHHPPHPPHPPH	HPHPHPHPHPPHPHPHPHPH	HHPHHPHPHHHHPHHPHPHH

*Looking for **optimal generation size** and **minimum number of generations** in a 2D-squared lattice*



	Some of t		
L	HP ≈ 1/2		
10	HPPHPPHPPH	H	
12	HPPPHPPHPPPH	HI	
14	HPPHPPHPPHPPPH	HPH	
16	HPPHPPHPPHPPHPPH	HPHPHPHPPHPHPHPH	HHPHHPHHPHHPHHPH
18	HPPHPPPHPPHPPHPPPH	HPHPHPHPHPPHPHPHPH	HPHHPHHPHHPHHPHHPH
20	HPPHPPHPPHHPPHPPHPPH	HPHPHPHPHPPHPHPHPHPH	HHPHHPHPHHHHPHHPHPHH

Concluding remarks

From experiments carried out we can conclude that:

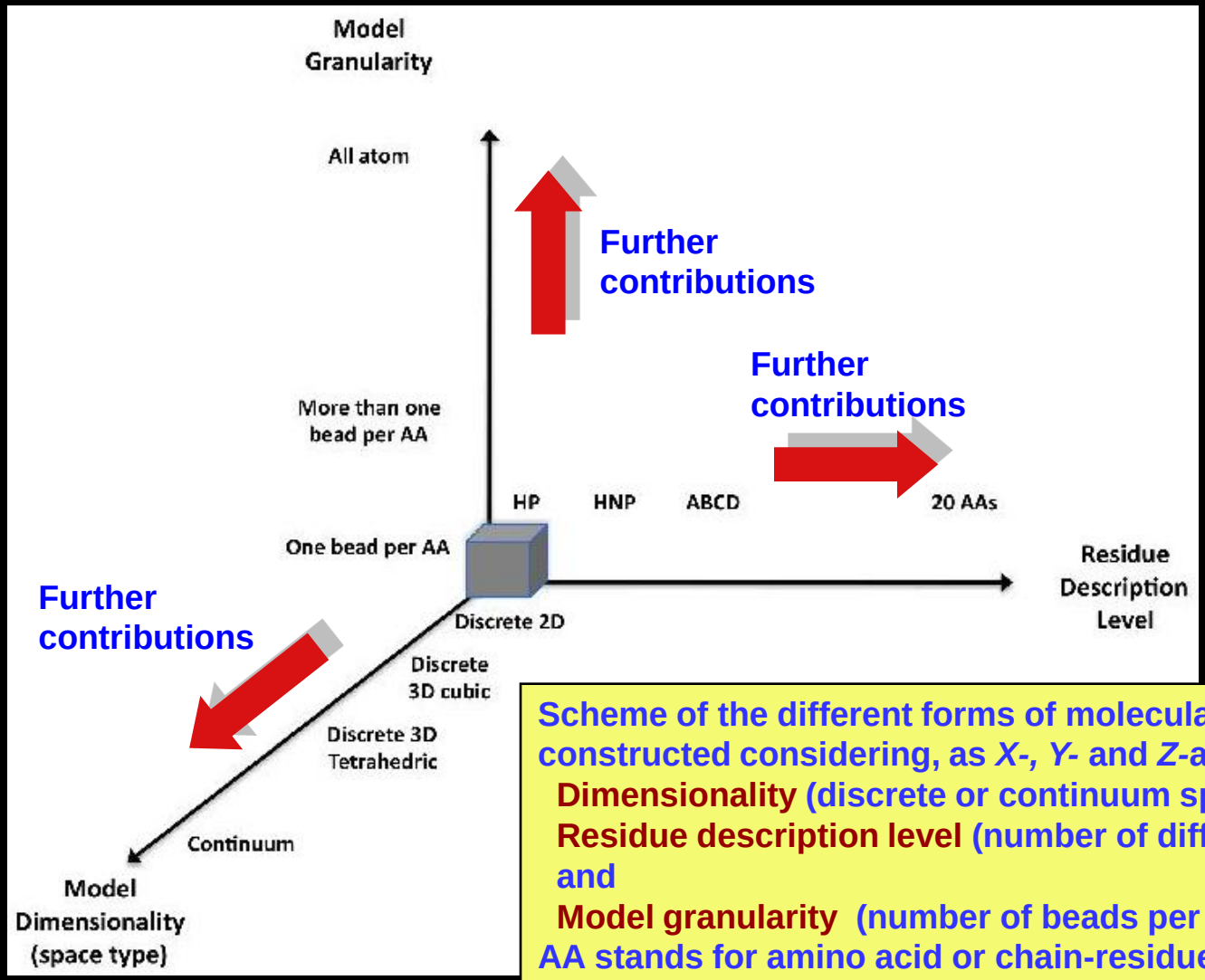
- Relatively modest experiments can yield a good amount of information to interact with the platform and to provide feedback for its adaptation to the particular issue to be explored.
- The bioinformatics platform is designed to incorporate any upgrades such as the ones suggested by the results of the various experiments.
- The platform is able to provide and integrate a very useful set of tools to the bioinformatics or biologist user to explore a great amount of possibilities, incorporating and updating an endless amount of indexes, correlations and subroutines, depending on the problem under study.

Concluding remarks

- The variety of tools will incrementally and opportunistically emerge depending on usage of the platform and knowledge of the biological problematic to be explored, these can be of physical, chemical, statistic and biological nature.
- Finally, the fine tuning of the gathered tools enables a particular accuracy of the solution of each tested problem depending on its complexity. In the sense that simpler problems will require less tools and complex problems will require more of them.

Concluding remarks

Description of the protein folding problem in the Evolution virtual laboratory



Scheme of the different forms of molecular folding representation constructed considering, as *X*-, *Y*- and *Z*-axes, the model:
Dimensionality (discrete or continuum space),
Residue description level (number of different amino-acid types), and
Model granularity (number of beads per monomer), respectively.
AA stands for amino acid or chain-residue.
H, N and P for hydrophobic, neutral and polar residues.

Concluding remarks

Description of the protein folding problem in the Evolution virtual laboratory