An Improved Immune Algorithm for Protein Folding

UNIVERSITY OF BIRMINGHAM

Eleanor Turpin, School of Computer Science

Introduction

This work is about solving the protein structure prediction (PSP) problem which is concerned with predicting the 3D native conformation of a protein from the corresponding 1D amino acid sequence. It is postulated that the native conformation corresponds to the global minima of the free energy; hence the PSP problem can be considered as an optimisation problem.

The clonal selection algorithm (Cutello *et al.*, 2007) is a state of art algorithm for this problem. In this work new, immune inspired, operators have been incorporated to improve the clonal selection algorithm's performance.

Clonal Selection Algorithm and the Biological Immune System

The clonal selection algorithm is an optimisation algorithm based upon the clonal selection principle in the biological immune system.

The clonal selection principle describes the response of cells from the adaptive immune system, called lymphocytes, to antigens. Lymphocytes are divided into T cells and B cells. When a B cell is exposed to an antigen it recognises it begins to proliferate (clone). The clones terminate as either antibody secreting plasma cells or B-memory cells that circulate throughout the host in case of re-infection. The clones undergo mutation of the gene region responsible for recognising antigens; this allows the immune system to adapt to recognise new antigens (de Castro & Timmis, 2002).

Immune Cell Diversity and the Mixed Strategy Operator

The biological immune system is made up of many types of cell that work both co-operatively and competitively. Examples of immune cells are macrophages, neutrophils, natural killer cells, phagocytes, plasma cells, TH_1 cells, TH_2 cells and CD8+ T cells all of which perform different functions.

The mixed strategy operator is based upon the cell diversity in the immune system. Which crossover or mutation technique is used to create offspring depends on a probability distribution $\underline{p} = (p_{hyp}, p_{macro}, p_{1pt}, p_{uni})$. The technique that creates the most offspring selected for the next generation is rewarded with a payoff that increases its probability of being selected.

The mixed strategy operator helps overcome the large energy barriers between low energy conformations of the HP-model proteins by allowing for large jumps in the search space, or 'tunnelling', between minima.

1. Initialise a random population of d solutions, $P^{(t=0)}$.
2. REPEAT
   i. Clone each solution dup times to produce a population $P^{(clo)}$.
   ii. Apply mixed strategy operator to $P^{(clo)}$.
   iii. Apply aging operator to $P^{(clo)}$ and $P^{(offspring)}$.
   iv. Select the d best solutions from $P^{(clo)} + P^{(offspring)}$ to form the next generation $P^{(t)}$.
   v. $t = t + 1$.
   vi. Update probability distribution $\underline{p}$.
3. UNTIL stopping criteria are met.

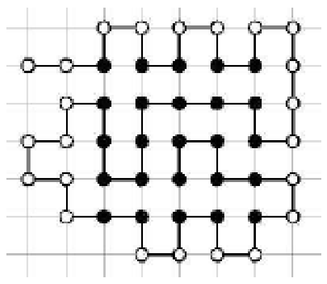
The HP Protein Model

This is the simplest model of protein-like structures and assumes that hydrophobic interactions are the dominant force in protein folding.

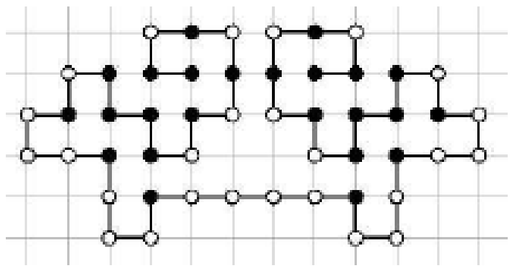
Free energy $E = \sum_{i \leq j} \varepsilon_{ij} \Delta_{ij}$

where $\varepsilon_{HH} = -1,\ \varepsilon_{HP} = \varepsilon_{PP} = 0$

and $\Delta_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are topological, but not sequential neighbours} \\ 0 & \text{otherwise} \end{cases}$



Above: Example of a native conformation of Sequence 5, E = -23 (Cox, 2007)



Left: Example of a native conformation of Sequence 6, E = -21, (Cox, 2007)

Immune Memory and the Archive Operator

Biological immune memory is acquired through:

- Presence of long-lived B-memory cells that exist in a resting state until a second encounter with an antigen
- Repeated exposure to antigens even in the absence of infection.
- Cross reactivity: the ability of the immune system to provide a secondary response to an antigen that it has not previously been exposed to, but which is similar to one that it has (de Castro & Timmis, 2002)

The archive operator is inspired by immune memory. A memory, or archive, of the best solutions is maintained. If a new solution is created that is similar too, but not better than, an archived solution, it is penalised. This is akin to memory cells that recognise antigens similar to ones encountered before rapidly proliferating; this proliferation can be considered as a payoff to the memory cell.

After each generation the archive is updated to remove any solutions that are too similar.

The archive operator helps overcome the large energy barriers between minima by encouraging population diversity and exploration of new areas of the search space.

```
FOR(j=0 to j=ARCHIVE-LENGTH)
    FOR(i=0 to i = N_c)
        IF H(r_i^offspring, r_j^archive) ≤ m :
            IF (f_i^offspring ≤ f_j^archive)  f_j^archive = f_j^archive + λ
            ELSE r_j^archive = r_i^offspring
    END FOR
END FOR
```

Benchmark Sequences

Instance	*l*	*Protein Sequence*	*E**
Sequence 1	20	$hphp_2h_2php_2hph_2p_2hph$	-9
Sequence 2	24	$h_2p_2(hp_2)_6h_2$	-9
Sequence 3	25	$p_2hp_2(h_2p_4)_3h_2$	-8
Sequence 4	36	$p_3h_2p_2h_2p_5h_7p_2h_2p_4h_2p_2hp_2$	-14
Sequence 5	48	$p_2h(p_2h_2)_2p_5h_{10}p_6(h_2p_2)_2hp_2h_5$	-23
Sequence 6	50	$h_2(ph)_3ph_4p(hp_3)_2hp_4(hp_3)_2hph_4(ph)_3ph_2$	-21
Sequence 9	20	$h_3p_2(hp)_2hp_2(hp)_2hp_2h$	-10
Sequence 13	18	$php_2hph_3ph_2ph_5$	-9
Sequence 14	18	$hphph_3p_3h_4p_2h_2$	-8
Sequence 15	18	$h_2p_5h_2p_3hp_3hp$	-4

l is the sequence length; *E** is the lowest known energy value

Results

	Mini-Project Results				New Results			
	Hypermutation & Macromutation		Hypermutation & Crossover		Mixed Strategy Operator		Mixed Strategy & Archive Operators	
	SR	AES	SR	AES	SR	AES	SR	AES
Sequence 1	100.00	33,313.43	100.00	**1,833.80**	100.00	6,478.27	100.00	5,647.47
Sequence 2	100.00	66,207.20	100.00	**5,082.12**	100.00	25,444.33	100.00	19,235.97
Sequence 3	100.00	564,214.37	100.00	80,546.17	100.00	**27,019.83**	100.00	64,618.90
Sequence 4	63.33	3,704,803.58	93.33	263,629.25	**100.00**	**425,302.53**	**100.00**	783,171.60
Sequence 5	3.33	2,222,925.00	0.00	0.00	80.00	944,266.58	**100.00**	**4,086.17**
Sequence 6	93.33	3,102,295.29	93.33	344,026.75	**100.00**	**93,461.97**	**100.00**	156,683.93
Sequence 9	100.00	15,502.13	100.00	3,749.73	100.00	5,098.90	100.00	**2,120.33**
Sequence 13	100.00	105,371.43	100.00	18,939.63	100.00	**16,012.83**	100.00	33,779.47
Sequence 14	100.00	63,019.47	100.00	**590.63**	100.00	15,983.10	100.00	18,914.23
Sequence 15	100.00	77,066.67	-	-	100.00	**25,751.90**	100.00	29,531.57

SR is the percentage of the repeat runs that found the lowest known energy value; AES is the average number energy function evaluations used to find the minima

Conclusions

The new operators improve the clonal selection algorithm's success of finding the lowest known energy and reduce the number of energy function evaluations used to find it. This is because the mixed strategy operator allows for large jumps that tunel between minima and the archive operator encourages diversity and exploration of the search space away from the current minima.

References

CUTELLO, V., NICOSIA, G., PAVONE, M., TIMMIS, J., 2007. An Immune Algorithm for Protein Structure Prediction on Lattice Models. *IEEE Transactions on Evolutionary Computation*, 11(1), pp.101-117.

DE CASTRO, L. N., TIMMIS, J., 2002. *Artificial Immune Systems: A New Computational Intelligence Approach*. London: Springer.

COX, G., 2007. *Protein Modelling: Protein Structure Prediction and Energy Landscape Analysis*. Thesis (phD). University of Birmingham

Acknowledgements: Project supervisor Jun He and Roy Johnston and Andrew Bennett of the School of Chemistry for help and advice; Giuseppe Nicosia of the University of Catania for kindly providing the code for the original clonal selection algorithm.