Evolution of 'Ligand-Diffusion Chreodes' on Protein-Surface Models: A Genetic-Algorithm Study

by **Sayed-Amir Marashi***[a])[b])[1]), **Mehdi Kargar**[b])[c])[1]), **Ali Katanforoush**[b])[d]), **Hassan Abolhassani**[b])[c]), and **Mehdi Sadeghi**[b])[d])[e])

[a]) Department of Biotechnology, University College of Science, University of Tehran, Enghelab Avenue, Tehran, Iran (phone/fax: +98-21-6649-1622; e-mail: marashie@khayam.ut.ac.ir)

[b]) School of Computer Science, Institute for Studies in Theoretical Physics and Mathematics (IPM), Niavaran Square, Tehran, Iran

[c]) Computer Engineering Department, Sharif University of Technology, Azadi Avenue, Tehran, Iran

[d]) Institute of Biochemistry and Biophysics, University of Tehran, Enghelab Avenue, Tehran, Iran

[e]) National Institute for Genetic Engineering and Biotechnology, Tehran-Karaj Highway, Tehran, Iran

Lattice models have been previously used to model ligand diffusion on protein surfaces. Using such models, it has been shown that the presence of pathways (or 'chreodes') of consecutive residues with certain properties can decrease the number of steps required for the arrival of a ligand at the active site. In this work, we show that, based on a genetic algorithm, ligand-diffusion pathways can evolve on a protein surface, when this surface is selected for shortening the travel length toward the active site. Biological implications of these results are discussed.

Introduction. – All biological macromolecules are surrounded by a layer of structured water (H_2O) molecules [1]. The state of H_2O molecules at protein surfaces is fundamental to protein structure, stability, dynamics, and function.

Lattice models have been extensively used to study a variety of chemical and biochemical interactions and reactions [2], including lateral-diffusion phenomena [3–6]. In recent years, lattice models have been used to demonstrate that ligand-diffusion paths or 'chreodes' can decrease the number of random steps that a ligand passes to arrive at its specific binding site [7–11]. These models generally assume *i*) that amino acids are non-randomly distributed on the protein surface; *ii*) that, over time, the number of H_2O molecules that are bound to a residue is a function of the amino acid hydropathy index (although the number of H_2O molecules can change); *iii*) that, when a ligand enters the hydration layer of a protein, there is never any back-diffusion to bulk H_2O; and *iv*) that a ligand can 'walk' on a protein surface until it reaches its specific binding site, the probability of its association to the adjacent residue being proportional to the number of H_2O molecules that are attached to that residue.

So far, using cellular-automata models, it has been shown that the existence of hydrophobic pathways or 'chreodes' on protein surfaces enhances the speed of ligand passage toward the active site. In this work, we demonstrate by means of a genetic

[1]) These authors contributed equally to the manuscript.

algorithm that hydrophobic residues tend to be arranged as pathways in our lattice model to minimize the number of steps that a ligand passes to arrive at the active site.

Experimental. – To eliminate the effect of outliers, we decided to use median instead of mean. We simulated the travel of a ligand toward an active site with the assumptions described in [11], with two differences: *1*) in contrast to previous works [7][9–11], we assumed that the number of H_2O molecules are constant, proportional to the hydropathical value of each cell. Therefore, our simulations only consider the 'most probable' states for the chreodes and it is constant. *2*) The observed travel lengths can be any integer value in the range of $[M, \infty]$, in which M is the minimum required steps for the arrival of the ligand at the active site. Therefore, the distribution of travels is skewed to right, and hence, very long travel lengths may be observed. It is necessary to remove these outliers before the calculation of the average travel length [11]. In the present work, median was used as the desired statistic to describe the properties of travel-length distributions. Median has the advantage that it is not influenced in the presence of outliers.

Here, we wanted to investigate whether a force for the minimization of travel length can result in the evolution of special amino acid patterns (*e.g.*, pathways) ending in the active site. For this purpose, we developed a genetic algorithm to change the amino acids on protein-surface models and to select for the best proteins (*i.e.*, with minimum travel length). The 'survivors' of this criterion are the parents of the next generation.

The pseudo-code of our program is illustrated in *Fig. 1.* In our program, for mutating a grid protein surface, a cell in the grid (*i.e.*, a residue on the protein surface) is randomly selected, and then its content is randomly changed to an integer in the range of 0 (most hydrophilic) to 9 (most hydrophobic). For performing a crossover, two distinct parents are selected, and then either half or a quarter of one is replaced by its counterpart region in the second matrix.

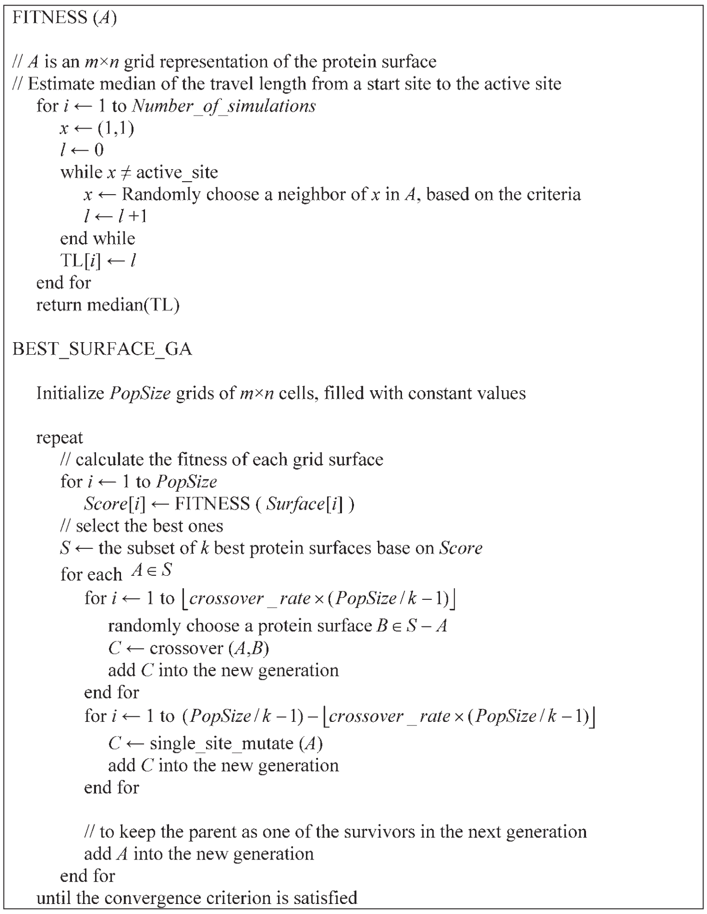
Construction of the offspring is done by a constant crossover rate in the range of [0,1]. The original parents ('survivors') are also present in the next generation. The remaining members of the new generation are then constructed by mutation. Running the genetic algorithm was continued until, for at least G generations, no improvement in the offspring scores was found.

In the simulations, two kinds of neighborhoods were possible. By default, we used the *Moore* neighborhood, in which it is assumed that all eight cells that surround a certain cell in a grid are their neighbors. We also used the *von Neumann* neighborhood in some of our simulations, which assumes that only four cells are in the neighborhood of a certain cell (*i.e.*, its up, down, left, and right adjacent cells).

Results and Discussion. – We modeled a protein surface with an $m \times n$ grid, in which each cell represents a residue on the protein surface. The details of pattern evolution on these grids are explained in the *Experimental* in detail. Briefly, these grids were randomly 'mutated', and the grids with minimum travel lengths were selected. The 'survived' grids were then used to construct the next generation, with replication and new random mutations. With a certain crossover rate, these 'survived' grids were allowed to exchange one, two, or three quarter of their surfaces.

Fig. 2 shows examples of the patterns evolved by our genetic algorithm. These patterns are evolved merely with the consideration of mutations, and without crossover (*i.e.*, crossover rate = 0). Note that these patterns match well with the 'chreodes' hypothesized to be present in previous studies [7][9–11].

Fig. 3 demonstrates an example of changing the average travel length during the generations. The standard deviation at the beginning is the highest, and generally decreases during consequent generations; however, this decrease is not a gradual trend, and, occasionally, the standard deviation increases in the next generations. Our simulation guarantees that some level of randomness is kept during the run, and this effect is observed because of this fact. In addition, during generations 7–9, changes in



```
FITNESS ($A$)

// $A$ is an $m \times n$ grid representation of the protein surface
// Estimate median of the travel length from a start site to the active site
    for $i \leftarrow 1$ to Number_of_simulations
        $x \leftarrow (1,1)$
        $l \leftarrow 0$
        while $x \neq$ active_site
            $x \leftarrow$ Randomly choose a neighbor of $x$ in $A$, based on the criteria
            $l \leftarrow l+1$
        end while
        TL[$i$] $\leftarrow l$
    end for
    return median(TL)

BEST_SURFACE_GA

    Initialize PopSize grids of $m \times n$ cells, filled with constant values

    repeat
        // calculate the fitness of each grid surface
        for $i \leftarrow 1$ to PopSize
            Score[$i$] $\leftarrow$ FITNESS ( Surface[$i$] )
        // select the best ones
        $S \leftarrow$ the subset of $k$ best protein surfaces base on Score
        for each $A \in S$
            for $i \leftarrow 1$ to $\lfloor crossover\_rate \times (PopSize/k - 1) \rfloor$
                randomly choose a protein surface $B \in S - A$
                $C \leftarrow$ crossover ($A$,$B$)
                add $C$ into the new generation
            end for
            for $i \leftarrow 1$ to $(PopSize/k - 1) - \lfloor crossover\_rate \times (PopSize/k - 1) \rfloor$
                $C \leftarrow$ single_site_mutate ($A$)
                add $C$ into the new generation
            end for

            // to keep the parent as one of the survivors in the next generation
            add $A$ into the new generation
        end for
    until the convergence criterion is satisfied
```

Fig. 1. *Genetic-algorithm* (GA) *pseudo-code used*. This algorithm aims at finding protein-surface models with patterns that minimize the travel length of a ligand over a protein surface.

travel length decreased, but at generation 10, a new decreasing trend started. Such a multi-step pattern is generally seen in our simulations (data not shown). This pattern presumably shows that the population has moved toward a local minimum; however, because of the mutations, it has the ability to overcome this local minimum and to be continuously improved.

Fig. 4 shows that, using the crossover method, the results look a bit more symmetrical, but are generally similar to those observed in *Fig. 2*. This suggests that the application of crossover (in addition to mutation) is not required for the improvement of the genetic-algorithm results.

Fig. 5 demonstrates the patterns evolved with the consideration of the *von Neumann* neighborhood, instead of the *Moore* neighborhood. We used 21 × 21 matrices to get more-obvious patterns. In fact, for 11 × 11 matrices, we had only accumulated amorphous mutations around the active site (data not shown).

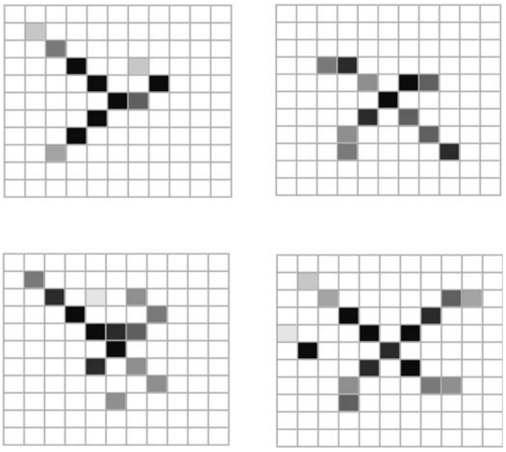


Fig. 2. *Examples of protein-surface representations evolved by our genetic algorithm without crossover*. In these simulations, the protein surface was modeled by an 11 × 11 grid of residues (initially filled with the most-hydrophilic residues), and the active site was at the center of the grid. The simulations were performed assuming the *Moore* neighborhood.

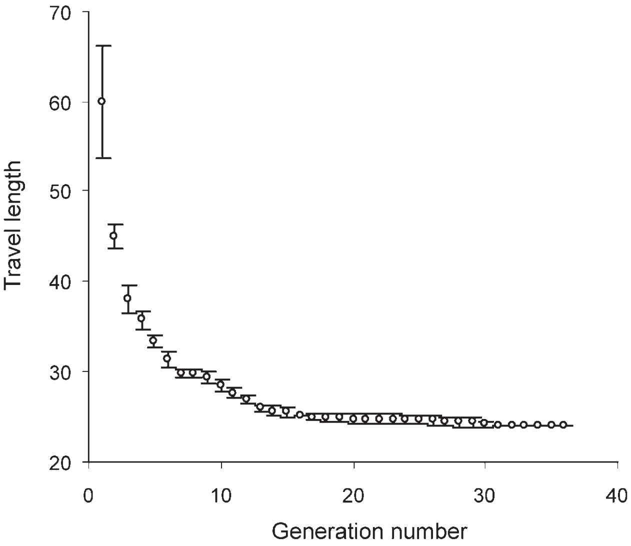


Fig. 3. *Example of travel-length evolution during generations*. In each generation of the experiments, 300 mutated 11 × 11 grids were constructed, and selected for minimizing the median of 300 simulated travel lengths. Each error bar is equal to the standard deviation of 300 travel lengths. See text for details.

Interestingly, we observed that, in most of the evolved pathways on the protein surfaces, more-hydrophobic residues are found around the active site. This is consistent with the assumption applied by *Kier* and co-workers in their modeling [9]. The

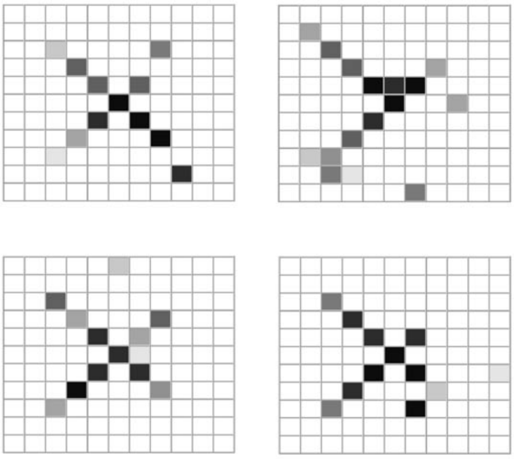


Fig. 4. *Examples of protein-surface representations evolved by our genetic algorithm, with a crossover rate of 0.25.* In these simulations, the protein surface was modeled by an 11 × 11 grid of residues (initially filled with the most-hydrophilic residues), and the active site was at the center of the grid. The simulations were performed assuming the *Moore* neighborhood.

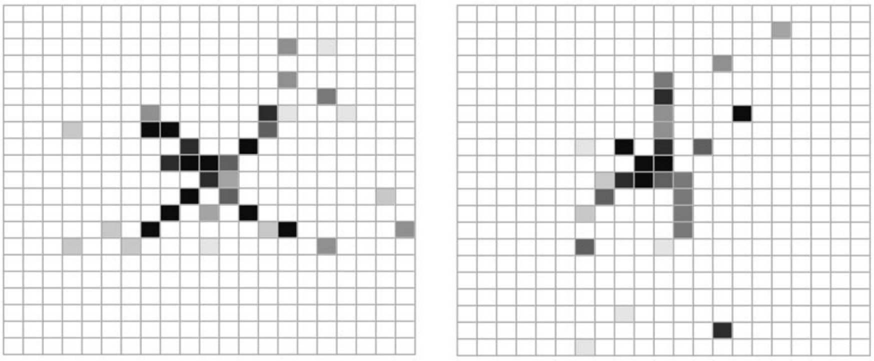


Fig. 5. *Examples of protein-surface representations evolved by our genetic algorithm, with a crossover rate of 0.25.* In these simulations, the protein surface was modeled by a 21 × 21 grid of residues (initially filled with the most-hydrophilic residues), and the active site was at the center of the grid. The simulations were performed assuming the *von Neumann* neighborhood.

existence of such a pattern suggests that not only the presence of such a pathway is important, but also the pathway performs best when it is 'steep' toward the active site.

Previously, it has been reported that, at least in some proteins, there are 'tunnels' that direct the movements of ligands headed for the enzyme active site [12][13]. We suggest that this notion should be revisited, considering the properties of the residues that construct the 'tunnel walls', as the composition of the walls may influence the structure of the H_2O molecules associated with the protein.

This research is, in part, supported by a grant from IPM (No. CS1385-1-02).

REFERENCES

[1] W. Qiu, Y. T. Kao, L. Zhang, Y. Yang, L. Wang, W. E. Stites, D. Zhong, A. H. Zewail, *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13979.

[2] K. Takahashi, S. N. Vel Arjunan, M. Tomita, *FEBS Lett.* **2005**, *579*, 1783.

[3] F. Drepper, I. Carlberg, B. Andersson, W. Haehnel, *Biochemistry* **1993**, *32*, 11915.

[4] M. J. Saxton, *Biophys. J.* **1987** *52*, 989.

[5] M. J. Saxton, *Biophys. J.* **1989**, *56*, 615.

[6] I. G. Tremmel, H. Kirchhoff, E. Weis and G. D. Farquhar, *Biochim. Biophys. Acta* **2003**, *1607*, 97.

[7] M. Ghaemi, N. Rezaei-Ghaleh, M.-N. Sarbolouki, *Lect. Notes Comput. Sci.* **2004**, *3305*, 719.

[8] L. B. Kier, *Am. Assoc. Nurse Anesth. J.* **2003**, *71*, 422.

[9] L. B. Kier, C. K. Cheng, B. Testa, *J. Theor. Biol.* **2001**, *215*, 415.

[10] L. B. Kier, C. K. Cheng, B. Testa, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 255.

[11] S.-A. Marashi, R. Behrouzi, *Biochem. Biophys. Res. Commun.* **2005**, *333*, 1.

[12] J. Liang, H. Edelsbrunner, C. Woodward, *Protein Sci.* **1998**, *7*, 1884.

[13] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, J. Damborský, *BMC Bioinformatics* **2006**, *7*, 316.

Received March 1, 2007