

Architecture of a VLSI chip for modelling Amino Acid Sequence in Proteins

S. Mitra S. Das P. Pal Chaudhuri
 Department of Computer Sc. & Engg.
 Indian Institute of Technology
 Kharagpur - 721 302, India
 e-mail : ppc@cse.iitkgp.ernet.in

S. Nandi
 Department of Computer Sc. & Engg.
 Indian Institute of Technology
 Guwahati - 781 001, India

Abstract

A Cellular Automata (CA) based model of amino acids which constitute different types protein is reported in this paper. A simulation engine is being developed based on this model to study protein behaviour.

1 Introduction

By the time we entered the last decade of this century, the fact that there is no substance so important as DNA(deoxyribonucleic acid) got well established. It is the storehouse of **biological information** of all living organisms on earth. It also provides the **molecular instruction** executed by different cells of a living being. Gene is a segment of ubiquitous DNA molecule that controls the function of all the cells in a living organism. Within the nucleus of each cell there are cromosomes, tangled tufts that contains DNA's long twisted strands. Those strands of DNA carry all the biochemical instructions necessary to produce and co-ordinate the components of living tissue. A DNA strand contains a long chain of four nitrogenous bases - Ademine(A), Guanine(G), Cytosine(C), and Thymine(T). A gene is a segment of DNA that contains the specific code necessary to produce a particular protein. Thus the genetic code consisting of a sequence of A's, G's, C's, and T's, encodes the information that controls the growth and activities of any living being through the synthesis of proteins - the three dimensional molecules that comprise the physical and chemical fabric of life.

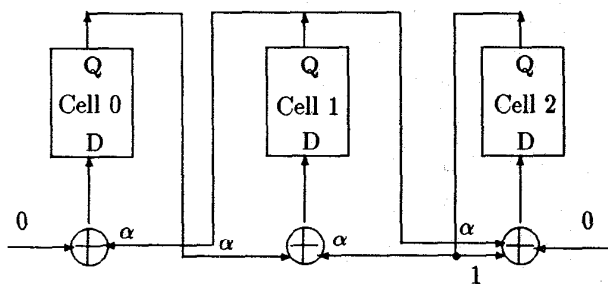
Proteins are large molecules made of small units linked into chains. These small units are amino acids. There are twenty different basic amino acids that can be joined to one another in various combinations to form a nearly infinite variety of protein chains. Out of four bases (A, G, C, T) [1], a group of three (codon) in the genetic code represents a particular amino acid. The DNA base sequence CGT, for example, stands for the amino acid **alanine**.

In the above background ,we are developing a simulation model to study the behaviour of the sequence of amino acids that constitute all different types of protein. This extended abstract reports only the robust model we have developed for study of amino acids.

The simulation engine is based on one dimensional ar-

ray of Cellular Automata (CA) cells. Each cell is capable of storing four values A(00), G(01), C(10),T(11) - that is instead of $GF(2)$ CA [2, 3] we shall work with $GF(2^2)$ CA. A CA with three cells having specific interconnection structure among them represent a specific amino acid. Such a CA is henceforth referred to as **Amino Acid CA (AACCA)**. Based on the analytical framework of $GF(2^2)$ algebra, AACCA machines are proposed to model the different amino acids. The specific CA structure is so chosen that the physico-chemical behaviour of the amino acid has necessary correspondence with the state transition behaviour of the corresponding AACCA. Next, a linear sequence of such AACCA cells are interconnected to represent a specific protein chain.

2 Modelling the Amino Acids with $GF(2^2)$ Cellular Automata



$$\text{Characteristic matrix} = \begin{bmatrix} 0 & \alpha & 0 \\ \alpha & 0 & \alpha \\ 0 & \alpha & 1 \end{bmatrix}$$

$$\text{Characteristic Poly} = \text{Minimal Poly} = x^3 + x^2 + \alpha$$

Figure 1: Structure & T matrix of a 3-cell $GF(2^2)$ CA

Let us consider a 3 cell $GF(2^2)$ CA (Figure 1). Each cell can store 0(00), 1(01), α (10) and α^2 (11). The next state of the linear CA depends on the weighted present state of the adjacent cells - i.e. the present

state is multiplied by the weight $w \in \{0, 1, \alpha, \alpha^2\}$. We consider CA's whose minimal polynomials are of the form $x^3 + ax^2 + bx + c$ where $a, b, c \in \{0, 1, \alpha, \alpha^2\}$. There are 64 such minimal polynomials. However, out of these 64, there are only 23 distinct minimal polynomials, each specifying a specific non-isomorphic state transition diagram. In other words, out of the state transition behaviour of 64 possible structures, only 23 are distinct - each differing in terms of cycles, cycle length, depth, etc. How each of these structures models a specific amino acid is next discussed.

The chemical formula of an amino-acid is of the form $R-NH_2-COOH$ i.e. different amino acids differ in the **R-group** also referred to as the side chain. As per the property of the R-group, amino acids may be classified into following categories [4]:

- (i) polar with positively charged R-group
- (ii) polar with negatively charged R-group
- (iii) polar but uncharged R-group
- (iv) nonpolar (hydrophobic) R-group

In the nature there exists 20 different amino acids as listed in Table 1. In addition there are 3 terminate codons [1]. Thus in total there are 23 different combinations of the triplet (of A, T, C, G) out of the $4^3 = 64$ combinations.

The above discussion points to the following inherent similarities in the proposed Amino Acid CA model (AACA) and the actual Amino Acid.

Table 1: Mapping of amino acids to AACA's

amino acid	Range of hydro.	partial specific vol.	AACA properties	
			max. cyc. len.	# of occurrence (cycle length)
Tyr	-3.4 to -2.3	0.703	1,2	1(1)
Trp		0.728		4(1)
Phe		0.766		4(1),6(2)
Ala	-1.8 to -0.5	0.732	3	1(1),1(3)
Val		0.831		1(1),5(3)
Ile		0.876		4(1),4(3)
Leu		0.876		4(1),20(3)
Pro	-1.8 to -0.5	0.748	4	4(1),6(2),12(4)
Cys	-1.8 to -0.5	0.630	5	1(1),3(5)
Met		0.739		4(1),12(5)
Ser	-0.5 to 0.3	0.596	6	1(1),1(3),2(6)
Thr		0.676		4(1),4(3),8(6)
His	-0.5 to 0.3	0.659	7	1(1),9(7)
Gly	-0.5 to 0.3	0.610	9	1(1),7(9)
Asn	-0.5 to 0.3	0.610	15	1(1),1(3),4(15)
Gln		0.667		1(1),1(3),3(5),3(15)
Asp	2.5	0.000	15	1(1),1(15)
Glu		0.106		4(1),4(15)
Arg	3.0	0.756	21,	1(1),3(21)
Lys		0.775		63

- 64 codons \Leftrightarrow 64 minimal polynomials of degree 3.
- 20 + 3 specific codons exist in nature \Leftrightarrow 23 distinct minimal polynomials of degree 3 with non-isomorphic state transition diagram possible.

Based on the study of the chemical structure and properties (aromaticity, sulphur content, etc) of Amino Acids, these have been divided into ten groups as noted in Table 1. We have mapped properties (hydrophobicity, partial specific volume) of 20 Amino Acids to the state transition behaviour of 20 distinct AACA based on the following criteria:

- Two groups differ in the maximum cycle length of the state transition diagram of the AACA's modelling the amino acids in those groups.
- Hydrophobicity of an amino acid is mapped to the length of the longest cycle of the state transition diagram of the corresponding AACA.
- Within each group, partial specific volume of an amino acid increases with the increase in total number of components in the state transition diagram of the corresponding AACA.

Based on this model we are developing a simulation engine to study the protein behaviour. It is almost impossible to perform simulation of the behaviour of the amino-acid sequences (protein) in software due to limitations in time and space. So, a simulation engine which can be realized in hardware is very much necessary. A VLSI chip for the simulation is being developed.

3 Conclusion

In this report we present the robust model of amino acids behaviour to study proteins in living organs.

References

- [1] T. E. Creighton, *PROTEINS Structures and Molecular Properties*. W H Freeman and Company, New York, 1993.
- [2] A. K. Das and P. P. Chaudhuri, "Efficient characterization of cellular automata," *Proc. IEE (Part E)*, vol. 137, pp. 81-87, January 1990.
- [3] M. Serra, T. Slater, J. C. Muzio, and D. M. Miller, "Analysis of one dimensional cellular automata and their aliasing probabilities," *IEEE Trans. on CAD*, vol. 9, pp. 767-778, July 1990.
- [4] A. L. Lehninger, *Principles of Biochemistry*. CBS Publishers & Distributors, Delhi, India, 1987.