RESEARCH PROJECT: TOWARD GALOIS THEORY OF PROTEIN-LIKE OBJECTS

NAOTO MORIKAWA

Contents

1. A quick introduction to Hetero number theory 2
1.1. Basic idea 2
1.2. Differential Geometry 3
1.3. Heterological Algebra 4
1.4. Examples 4
1.4.1. Natural numbers 4
1.4.2. Rhombic dodecahedron 5
1.4.3. DNA 6
1.4.4. Protein 6
2. Galois theory of hetero numbers 8
2.1. Algebraic equations of petals 8
2.2. Symmetry group of petals 8
2.3. Shape Description Language 9
3. Study plan 10
References 10

The ultimate aim of this project is to describe proteins w.r.t. their functions. Recall that proteins usually form complexes, either stable or transient, to perform their jobs ([1]). For example, nearly ten thousands of protein complexes are registered with the Protein Quaternary Structure (PQS) database (http://pqs.ebi.ac.uk/pqs-doc.shtml). That is, protein-protein interactions are key determinants of protein function. And we will try to describe proteins w.r.t. their partners, with which they form complexes.

In this project we consider, for example, the following questions:

- Are proteins uniquely determined by the protein-protein interaction map of an organism?
- Are there any kind of symmetry among protein-protein interactions?

We will construct a mathematical foundation to deal with them.

Moreover the structure-function paradigm of proteins implies a new kind of functional language, whose semantics is given by shapes. And we expect a "Protein Description Language" based on it to express specification of proteins, i.e., the protein-protein interaction map.

Date: August 31, 2003.
(rev.) October 31, 2003.

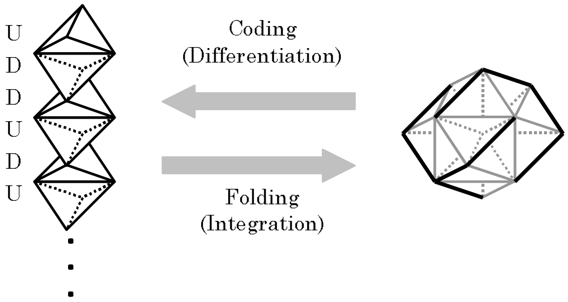


Figure 1. Analysis of hetero numbers.

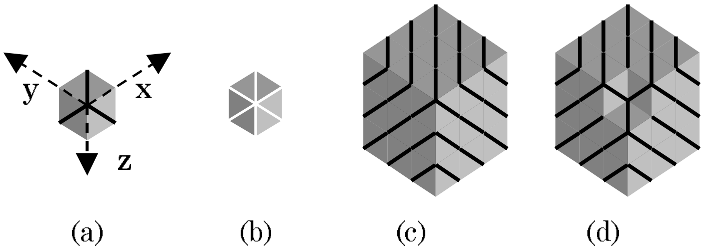


Figure 2. Unit cube and its drawings in $\mathbb{R}^3$

1. A quick introduction to Hetero number theory

We use rulers to measure length of objects and weights in scales for weighing objects. The set of "hetero numbers" is a new system of units for measuring shape of objects such as proteins. The features of the system are the correspondences between

(1) genetic code and the second derivative,
(2) protein folding and integration,
(3) protein-protein interaction and addition

(Fig.1). Moreover the system gives an example of "additively higher dimensional" extension of natural numbers.

1.1. **Basic idea.** We explain the basic idea of "hetero numbers" in the case of dimension two. Consider a unit cube in the three-dimensional Euclidean space $\mathbb{R}^3$ whose vertices are, say, given by $v_1 = (0,0,0)$, $v_x = (1,0,0)$, $v_y = (0,1,0)$, $v_{xy} = (1,1,0)$, $v_z = (0,0,1)$, $v_{xz} = (1,0,1)$, $v_{yz} = (0,1,1)$ and $v_{xyz} = (1,1,1)$. And draw lines $\overline{v_1 v_{xy}}$, $\overline{v_1 v_{yz}}$ and $\overline{v_1 v_{xz}}$ (Fig.2(a)).

Then each of the three upper faces is divided into two slant triangle tiles by the lines. For example, triangles $v_1 v_x v_{xy}$ and $v_1 v_y v_{xy}$ for the face $v_1 v_x v_{xy} v_y$. By projecting the faces into the hypersurface $x + y + z = 0$, we obtain a division of a hexagon by six triangle tiles (Fig.2(b)).

Piling up these unit cubes in the direction from v_{xyz} to v_1, we obtain a drawing made up of the lines (Fig.2 (c) and (d)). Note that its peaks uniquely determine the drawings. For example, the drawing of Fig.2(c) is determined by $(0,0,0)$ and Fig.2(d) by its three peaks $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$. We denote a set of peaks

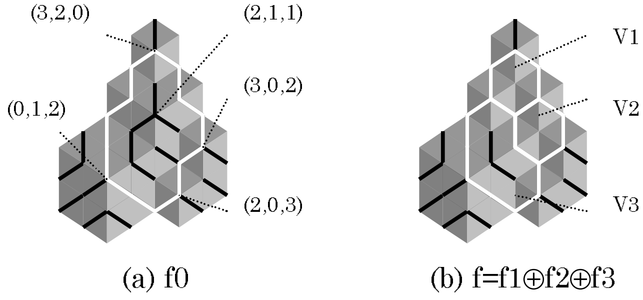


Figure 3. Affine hetero numbers

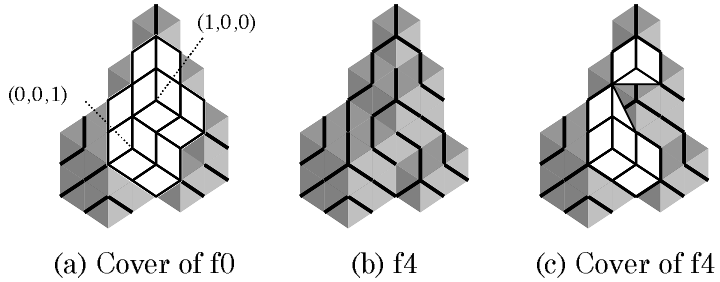


Figure 4. Cover of peaks

by a polynomial, where a term $x^l y^m z^n$ corresponds to a peak $(l, m, n) \in \mathbb{Z}^3$. That is, Fig.2(c) is determined by $1 = x^0 y^0 z^0$ and Fig.2(d) by $x + y + z$.

What concerns us is the case when a drawing gives closed orbits of tiles. For example, a drawing determined by a polynomial

$$f_0 = x^3y^2 + x^3z^2 + x^2z^3 + yz^2 + x^2yz$$

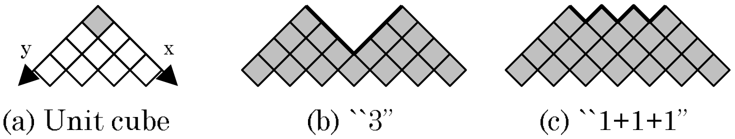
defines a closed orbit (Fig.3(a)). We call the set of all closed orbits *two-dimensional affine hetero numbers* and denote it by $\mathbf{AHN}^2$.

An affine hetero number is characterized by its "cover" (the least upper bound of its peaks). For example, the cover of f_0 is $x + z$ (Fig.4(a)). Suppose that a set of peaks are given and consider all slant triangle tiles of the drawing determined by them. If these tiles are either under or above the cover as in Fig.4(a), the peaks define an affine number.

On the other hand, if there exists a tile which intersects the cover as in Fig.4(c), the peaks do not define any affine numbers. For example, if we remove a term x^3z^2 from f_0, they define no longer any affine numbers (Fig.4(b)). For detailed discussion, see [3].

1.2. **Differential Geometry.** Orbits are uniquely determined by the gradient of its slant tiles, i.e., up (U) and down (D). Consider again the closed orbit defined by f_0 (Fig.3(a)), which consists of a chain of twenty-six triangle tiles. Moving clockwise from the tile $(3, 2, 0)(3, 2, 1)(4, 2, 1)$, gradients along the orbit are given by

$$D - U - U - D - D - U - D - U - \cdots - D - U.$$



(a) Unit cube (b) ``3'' (c) ``1+1+1''

FIGURE 5. 1-dim. affine hetero numbers

The string of two letters, D and U, gives a "genetic coding" of the shape covered by the orbit.

1.3. **Heterological Algebra.** Removing a cube at the peak $(2, 1, 1)$, we obtain another drawing determined by

$$f = x^3y^2 + x^3yz + x^3z^2 + x^2z^3 + yz^2 + x^2y^2z.$$

It defines three closed orbits $V1$, $V2$ and $V3$ (Fig.3(b)) determined by

$$\begin{aligned} f_1 &= x^3y^2 + x^3yz + x^2y^2z, \\ f_2 &= x^3yz + x^3z^2 + x^2yz^2, \\ f_3 &= x^2z^3 + yz^2 + x^2y^2z. \end{aligned}$$

We denote the relation among f, f_1, f_2 and f_3 as addition, i.e.,

$$f = f_1 \oplus f_2 \oplus f_3.$$

which is the "additive" prime factoring of the affine hetero number f.

By putting a cube on the drawing, f fuses into a single orbit f_0, i.e.,

$$f_0 = (f_1 \oplus f_2 \oplus f_3) * x^2yz,$$

which defines action of terms on affine hetero numbers.

1.4. **Examples.**

1.4.1. *Natural numbers.* A natural number gives an example of one-dimensional affine hetero numbers. We define an embedding of $\mathbb{N}$ into $\mathbf{AHN}^1$ by

$$k \mapsto x^k + y^k$$

(Fig.5(b)). Then we have

$$(x^k + y^k) * (\sum_{0<i<k} x^iy^{k-i-1}) = \oplus_{0<j\leq k}(x^jy^{k-j} + x^{j-1}y^{k-j+1}).$$

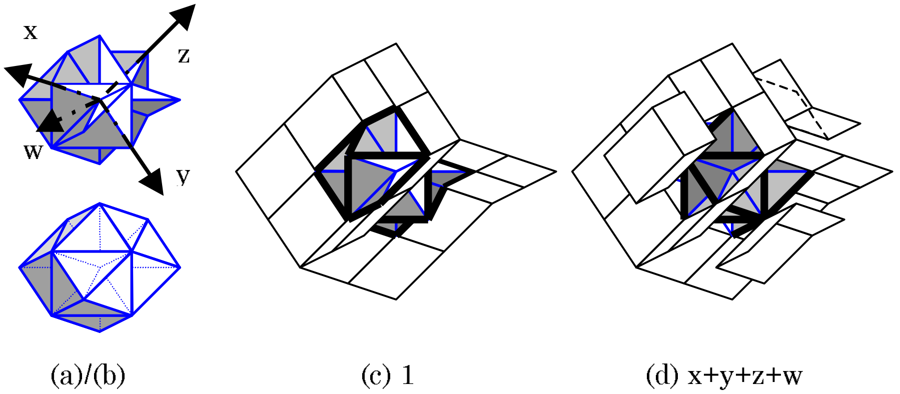
The left hand side of the equation corresponds to a natural number "k" and the right hand side corresponds to addition "$1 + 1 + \cdots + 1$ (k times)" (Fig.5(c)). The genetic code of "k" is

$$D - D - \cdots - D - U - U - \cdots - U(k\ D\text{s followed by } k\ U\text{s}).$$

On the other hand, the genetic code of "$1 + 1 + \cdots + 1$(k times)" is

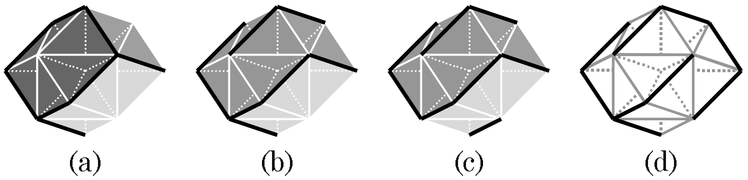
$$D - U - D - U - \cdots - D - U(k \text{ repetitions of } D - U- \ ).$$

We can also embed $\mathbb{N}$ into higher dimensional affine hetero numbers similarly.



(a)/(b) (c) 1 (d) x+y+z+w

FIGURE 6. Unit cube and its drawings in $\mathbb{R}^4$



(a) (b) (c) (d)

FIGURE 7. Rhombic dodecahedron

1.4.2. *Rhombic dodecahedron.* Next we consider three-dimensional affine hetero numbers. Note that a unit cube in $\mathbb{R}^4$ is projected onto a rhombic dodecahedron in $\mathbb{R}^3$ by a mapping defined by $(1,0,0,0) \mapsto (1,0,0)$, $(0,1,0,0) \mapsto (0,1,0)$, $(0,0,1,0) \mapsto (0,0,1)$ and $(0,0,0,1) \mapsto (-1,-1,-1)$.

Consider a unit cube in the four-dimensional Euclidean space $\mathbb{R}^4$ whose vertices are, say, given by $v_1 = (0,0,0,0)$, $v_x = (1,0,0,0)$, $v_y = (0,1,0,0)$, $v_z = (0,0,1,0)$, $v_w = (0,0,0,1)$, $v_{xy} = (1,1,0,0)$, $v_{xz} = (1,0,1,0)$, $\cdots$, $v_{xyzw} = (1,1,1,1)$.

Then each of the four upper faces, that is, the faces specified by $x = 0$, $y = 0$, $z = 0$ or $w = 0$, is divided into six tetrahedron tiles (Fig.6(a)). For example, six tetrahedrons $v_1 v_x v_{xy} v_{xyz}$, $v_1 v_x v_{xz} v_{xyz}$, $v_1 v_y v_{xy} v_{xyz}$, $v_1 v_y v_{yz} v_{xyz}$, $v_1 v_z v_{yz} v_{xzy}$ and $v_1 v_z v_{xz} v_{xyz}$ for the face specified by $z = 0$ (divided by the six white triangle walls in Fig.6(a)). And "up (U)" (resp. "down (D)") at the tetrahedrons corresponds to the direction from v_{xyz} to v_1 (resp. from v_1 to v_{xyz}). In particular, by projecting the faces into the hypersurface $x + y + z + w = 0$, we obtain a division of a rhombic dodecahedron by twenty-four tetrahedron tiles (Fig.6(b)).

Piling up unit cubes in the direction from v_{xyzw} to v_1, we obtain a drawing made up of the chains of tetrahedrons (Fig.6 (c) and (d)). Note that tetrahedrons are connected only in the direction of U or D. Fig.6(c) is determined by a polynomial 1 and Fig.6(d) by $x + y + z + w$. In some drawings, we obtain a decomposition of a rhombic dodecahedron into chains of tetrahedron tiles (Fig.7).

The chain of tetrahedrons satisfies the following conditions:

(1) Each tetrahedron consists of four short edges and two long edges, where the ratio of the length is $\sqrt{3}/2$.
(2) Tetrahedrons are connected via long edges and rotate around the edges.

Any shape represented by a polynomial is obtained by folding some chains of adequate length according to their "U/D" codes and holding them together.

Fig.7 corresponds to the following equations:

(a) $f = g_0 \oplus g_1 \oplus g_2 \oplus g_3$
(b) $f * xz = g_{10} \oplus g_{11} \oplus g_{12}$,
(c) $f * (xz + yz) = g_{20} \oplus g_{21}$,
(d) $f * (xz + yz + wz) = g_{30}$,

where $f = xyz + xyw + xzw + yzw$, $g_0 = xyz + xyw + xzw$, $g_1 = xyz + xyw + yzw$, $g_2 = xyz + xzw + yzw$, $g_3 = xyw + xzw + yzw$ $g_{10} = xyw + yzw + xz$, $g_{11} = xyz + xyw + yzw$, $g_{12} = xyw + xzw + yzw$, $g_{20} = xyw + yzw + xz$, $g_{21} = xyz + xzw + yw$ and $g_{30} = xyw + xz + yz + wz$.

The gradients of tiles along the orbit of g_{30} is

$$\begin{aligned} &U - D - D - U - D - U - U - D \\ &\qquad - U - D - D - U - D - U - U - D \\ &\qquad - U - D - D - U - D - U - U - D, \end{aligned}$$

which gives the genetic code of a rhombic dodecahedron (Fig.1).

1.4.3. *DNA.* An affine hetero number is defined by a single polynomial, that is , "affine". On the other hand, a DNA molecule is approximated by patching affine orbits together as in the case of "manifold". The correspondence between a DNA and a chain of tetrahedron tiles is given by

$$\text{one nucleotide} \Longleftrightarrow \text{one tetrahedron tile.}$$

Using the approximation, we obtain a polynomial representation of helix, which has twelve tiles per turn (Fig.8(b)):

$$\begin{aligned} p(\text{helix}) = \{&(1 + x^2y/w, [0,7]), \\ &(xy + x^2z^2 + x^2y/w, [2,13]), \\ &(x^2yz + x^2z^2 + x^4yz^2/w, [8,19]), \\ &(x^3yz^2 + x^4z^4 + x^4yz^2/w, [14,25]), \\ &(x^4yz^3 + x^4z^4 + x^6yz^4/w, [20,29])\}, \end{aligned}$$

where $(f, [l, m])$ means that the part from the l-th tile to the m-th tile is represented by f.

The genetic code of the helix (from top to bottom) is

$$\begin{aligned} &D - D - D - D - U - U - D - D - D - D \\ &\qquad - U - U - D - D - D - D - U - U - D - D \\ &\qquad - D - D - U - U - D - D - D - D - U - U. \end{aligned}$$

Two copies of the helix form a DNA-like double-helix as shown in Fig.8(a). In the figures arrows indicate the direction of "down (D)".

1.4.4. *Protein.* A protein molecule is also approximated by a set of affine orbits. The correspondence between a protein and a chain of tetrahedron tiles is given by

$$\text{one amino-acid} \Longleftrightarrow \text{three tetrahedron tiles.}$$

Note that the coding is consistent with the actual genetic code, where a single amino-acid is coded by three nucleotides (codon). Therefore the actual genetic

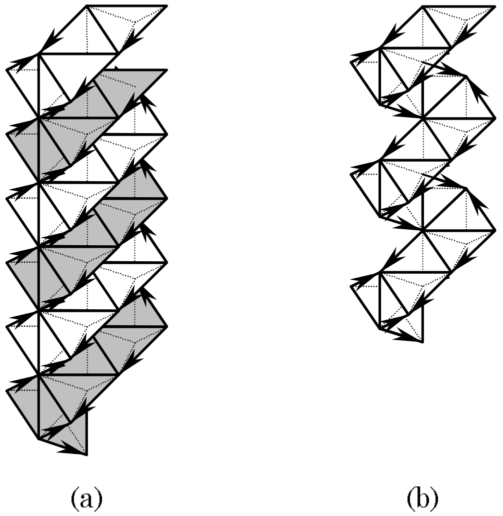


Figure 8. DNA

code of a protein is encoded into a string of two letters, U and D. In particular, the twenty kinds of amino-acids are encoded into eight kinds of letters. (In the following, for the convenience of the program used, a bond between amino-acids is corresponded to three tiles.)

Using the approximation, we obtain a polynomial representation of the chain A of 2HIU(Insulin, human):

$$\begin{aligned} p(\text{2HIU-A}) = \{ & (x/w + 1/(yz^2) + 1/(xz^2), [0,12]), \\ & (1/(zw) + 1/(yz^2) + 1/(xz^2), [5,16]), \\ & (1/(zw) + 1/(xyz^3) + 1/(x^2z^3) + y/(xzw), [11,27]), \\ & (1/(x^2zw) + 1/(xyz^3) + 1/(x^2z^3) + yz/w^2, [14,40]), \\ & (y^2z/(xw) + y/w + yz/w^2, [34,43]), \\ & (y^2z/(xw) + 1/w^2 + z/w^3, [38,49]), \\ & (z/(xw^4) + 1/w^2 + z/(yw^4), [44,55]), \\ & (z/(xw^4) + 1/(y^2w^4), [50,59])\}. \end{aligned}$$

('2HIU' is the ID used to retrieve data from the Protein Data Bank (PDB).)

The "D/U code" of the protein is

$$\begin{aligned} & D-U-D-U-U-U-U-D-D-U \\ & \quad -U-D-D-U-D-U-U-U-U-D \\ & \quad -D-U-U-D-D-D-D-U-U-U \\ & \quad -D-D-D-D-D-D-U-U-D-D \\ & \quad -U-D-U-U-U-U-D-D-U-U \\ & \quad -U-U-D-D-U-U-U-U-D-D. \end{aligned}$$

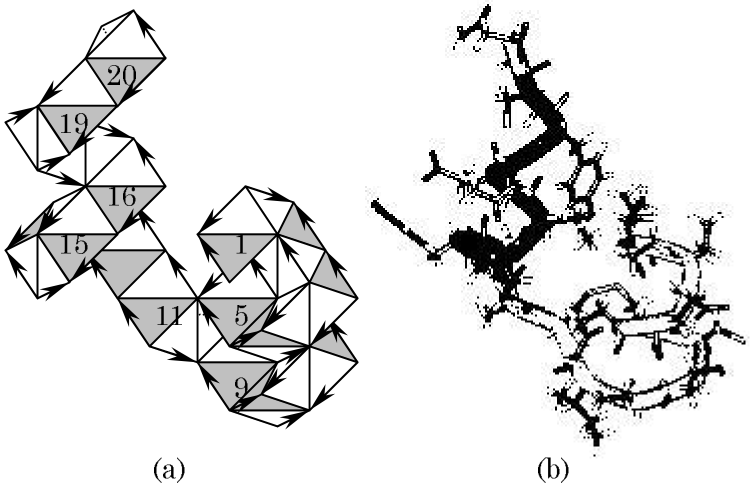


FIGURE 9. Protein(2HIU chain A). The figure (b) is prepared using **WebLab Viewer** (Molecular Simulations Inc.).

2. GALOIS THEORY OF HETERO NUMBERS

2.1. **Algebraic equations of petals.** Let's consider the closed orbit VA defined by a polynomial $p(VA) = xy+xz+yz$ and closed orbits Ai $(0 \leq i < 7)$ surrounding it (Fig.10(a)). We call these surrounding loops the *petals* of VA. Among VA and its petals, there are eight algebraic equations (and their combinations). For example, Fig.11 illustrates the following equations:

(a) $(p(VA) \oplus (\oplus_{1 \leq i \leq 6} p(Ai))) * 1 = p(A11) \oplus p(A12) \oplus p(A13)$,
(b) $(p(VA) \oplus (\oplus_{1 \leq i \leq 6} p(Ai))) * (x+y+z) = p(A21)$,
(c) $(p(VA) \oplus p(A5) \oplus p(A6)) * y = p(A31)$,
(d) $(p(VA) \oplus p(A4) \oplus p(A5)) * (-yz) = p(A41)$,
(e) $(p(VA) \oplus (\oplus_{1 \leq i \leq 6} p(Ai))) * (-xy - yz - xz) = p(A51) \oplus p(A52) \oplus p(A53)$

where $p(Ai)$ denotes the polynomial representation of the closed orbit Ai.

This observation leads us to expect the following correspondence (modulo some equivalence relation).

Problem 1.

An affine hetero number

$\overset{?}{\Leftrightarrow}$ A set of equations among the orbit and its petals.

That is, given a set of equations, we consider whether there exist any affine hetero numbers that satisfy them or not. Note that, to consider the problem, we need some language in which we express and solve the equations.

2.2. **Symmetry group of petals.** Because of the symmetry of VA, the equations are invariant under some permutation of the petals. In this case, they are invariant under any rotations and any reflections among Ai $(0 \leq i < 7)$.

Next consider a less symmetric orbit, i.e., the closed orbit VB defined by $p(VB) = x^2y^2z + x^2z^3 + yz^2$ and its petals Bi $(0 \leq i < 8)$ (Fig.10(b)). Then we have eleven

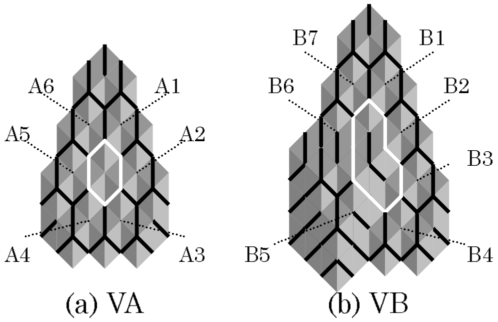


Figure 10. Affine hetero number and its petals

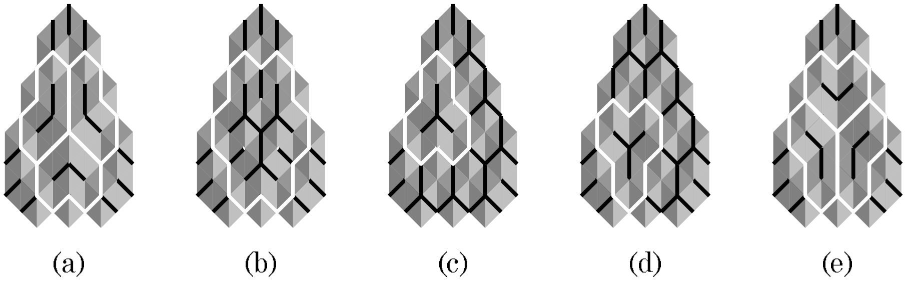


Figure 11. Algebraic equations of VA

equations (and their combinations) among them. For example, Fig.12 illustrates the following equations:

(a) $(p(VB) \oplus (\oplus_{1 \le i \le 7} p(Bi))) * (xyz + xz^2) = p(B11) \oplus p(B12) \oplus p(B13) \oplus p(B14)$,
(b) $(p(VB) \oplus (\oplus_{1 \le i \le 7} p(Bi))) * (x^2yz + xy^2z + x^2z^2 + xz^3) = p(B21) \oplus p(B22)$,
(c) $(p(VB) \oplus p(B6) \oplus p(B7)) * xy^2z = p(B31)$,
(d) $(p(VB) \oplus p(B5) \oplus p(B6)) * (-yz^2 + xyz^2) = p(B41)$,
(e) $(p(VB) \oplus (\oplus_{1 \le i \le 7} p(Bi))) * (-x^2y^2z - x^2z^3 - yz^2) = p(B51) \oplus p(B52)$.

Now they are not invariant under any rotations among Bi $(0 \le i < 8)$. This observation leads us to expect the following correspondence (modulo some equivalence relation).

Problem 2.

$$\text{Symmetry of an affine hetero number}$$
$$\overset{?}{\Leftrightarrow} \text{A set of permutations of its petals.}$$

And we regard the set of permutations as the "Galois group" of an affine hetero number.

2.3. **Shape Description Language.** If we have a programming language in which we can express and solve the equations efficiently, it would facilitate calculation of hetero numbers very much. Then a set of equations will become a computer program. And its solution is nothing but the "semantics" of the program.

Note that algebraic equations can be regarded as "syntax of terms" of the language. For example, $(p(VA) \oplus p(A5) \oplus p(A6)) * y = p(A31)$ (Fig.11(c)) implies that

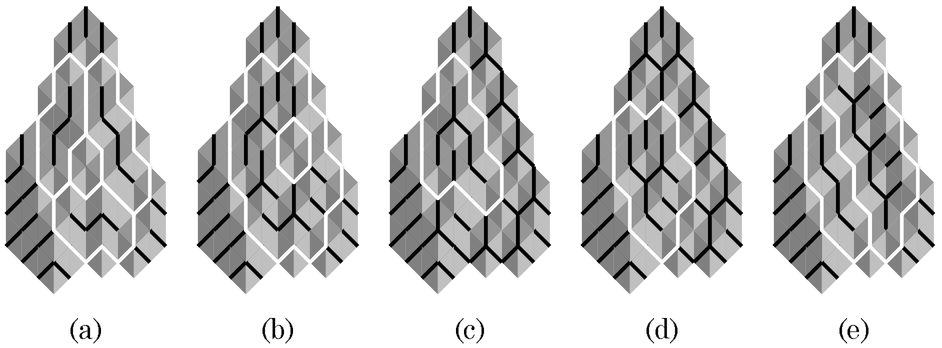


FIGURE 12. Algebraic equations of VB

$A31$ is a concatenation of VA, $A5$ and $A6$, which gives a two-dimensional extension of concatenation of lists such as "cons" in LISP.

This observation motivates the following.

Problem 3.

Design an higher dimensional extension of LISP (or λ-calculus)

to describe and manipulate hetero numbers.

3. Study plan

Purpose:

- To define the set of equations which determines an affine hetero number uniquely and establish the way to solve it.
- To give a specification of an affine hetero number using the symmetry of the equations which define it.

Strategy:

- By imitating sheaf theory (topos theory) as far as possible to clarify the difference between homology and "heterology". We define an embedding Δ of natural numbers into affine hetero numbers. Then the category of affine hetero numbers can be regarded as sheafs on $\{\Delta(k) : k \in \mathbb{N}\} \subset \mathbf{AHN}^m$. (Cf. The category $\mathbf{B}G$ of continuous G-set. ([2]))
- By imitating λ-calculus (or LISP) as far as possible to make the implementation of the language readily. As explained above. we regard addition as a higher dimensional extension of concatenation of lists.

References

1. A.Abbott, news feature The society of proteins, Nature 417, 894-896(2002).
2. S.Mac Lane, I.Moerdijk, Sheaves in Geometry and Logic, Springer, 1992.
3. N.Morikawa, *Polynomial representation of DNA and proteins*, 2002 (manuscript).

E-mail address: `nmorika@f3.dion.ne.jp`