FULL MATCHING AND ISOMORPHISM OF NUCLEOTIDE CHAINS

IMRE BÁLINT*, ANDRÁS KOCSOR, IMRE CSERPÁN, AND KATALIN FODOR

Abstract. The structure of DNA/RNA chains is determined at several different levels. Although the primary sequence is fundamental in determining further properties, the higher level organization may have a backward influence. Beyond the physical structure of the chain, an abstract 'shape' can also be defined. This is described in terms of relationships among the building blocks. We aim to show, that matching sequences are not exclusive representatives of relationships among nucleotide chains. Our approach is based on lattice theory, since lattice theory especially fits the problem. Beside the qualitative recognition of (algebraic) similarity, we also give a measure. This allows the structured quantification of the 'similarity distance' of two chains. Such a decomposition provides the possibility to recognize relationships between biological phenomena and deep-lying structural similarities.

1. Introduction

The immense development of automatic DNA/RNA-sequencing resulted in huge libraries of known genetic codes. The information content of these data banks is much larger, than exploited today and appropriate mathematical tools will open the way to the mining of these rich sources [1-7]. Such studies may enlighten relations between deep-rooted structural properties of DNA/RNA-chains and some higher-level properties of a living organism.

The aim is to define a 'generalized shape' for every DNA/RNA sequence and a way to their comprehensive characterization [8-11]. The mathematical definition of shape is not unique, it can be described in various ways emphazising different characteristics of the object. This is the reason that some branches of universal algebra, topology or algebraic topology [12-19] can serve as suitable (but not equivalent) tools of the 'shape-description' problem. Our aim is to define those structural similarities, which are beyond the simple matching. The structures of nucleotide chains will be studied by tools of lattice theory, because lattice algebra especially fits the problem. The relationship of subsequences allows a fine meshed characterization of the chain with a similarity measure, which quantifies the basically qualitative similarity concept of lattice theory.

The report is structured as follows: the next section is devoted to technicalities and the basic lattices are introduced with congruence relations. Subsequently the various isomorphism concepts are defined, then the measures with their carrier lattices and finally some simple illustrations close the paper. To make the reading easier, the proofs of assertions are collected

*The first author acknowledges the financial support of the Hungarian Science Fund (Grant No. OTKA T4450).

into a separate section (Appendix 1.), while some background material on lattice theory is given in Appendix 2..

2. The primary lattice structure of nucleotide chains

A nucleotide chain will be considered exclusively as a (single) sequence of letters,

$$AAACUAUUUUUAAAUAUGUUUUGAAAACAUGUUUU... \tag{1}$$
$$...GAGGUAACUCGGUAGUUUUCCA$$

as our example the 54-element $tRNA^{Ser}$ of the *C. elegans* [3]. Neither the double-stranded structure of DNA, nor the conformation or some quantum chemical parameter will be taken into account. All these properties are implicitly supposed to be determined by the primary nucleotide sequence. We shall deal with the internal relationships of the chain by defining a lattice algebra over the set of subsequences.

A word is an ensemble of juxtaposed letters chosen from a fixed alphabet, which form a connected string. In our case the words are given over alphabet $\{A, G, C, T\}$, which is mapped bijectively onto another (arbitrarily chosen) integer alphabet $\psi : \{A, G, C, T\} \longrightarrow \{1, 2, 3, 4\}$, where $\psi(A) = 1, ..., \psi(T) = 4$. The actual choice of mapping does not restrict the generality of the discussion, however the comparison of different chain molecules is only possible with identical mapping functions. The set of words over the nucleotide alphabet is C_{ch} ('nucleotide chains'), that over the integer alphabet is C_{iw} ('integer words'). The bijective map $C_{ch} \longleftrightarrow_{\psi} C_{iw}$ uniquely associates nucleotide sequences and words of integers. The elements of C_{iw} are k-words $S^k = s_1...s_k,\ k \in \mathbb{N}^+$, where $s_1, ..., s_k$ are elements of the integer alphabet and the whole word is read in left-to-right direction coinciding with the $3' - 5'$-direction of the underlying nucleotide chain. The elements of C_{iw} are distinguished by a subscript $S_r,\ r \in W$. If the length of a word is stressed, a superscript appears S^n_r. If $S^n = s_1...s_n$ is a general n-word, $S^n(k, l) = s_k...s_l,\ (1 \leq k \leq l \leq n)$ is a connected sub-sequence, i.e. a sub-word.

Definition 1. *Operator* $\tau : C_{iw} \longrightarrow \mathcal{P}(C_{iw})$,

$$\tau(S^n) = \{\ S^n(k, l)\ :\ 1 \leq k \leq l \leq n\ \},\quad S^n = s_1...s_n \in C_{iw},\ n \in \mathbb{N}^+ \tag{2}$$

associates every word with the set of its connected sub-words.

Definition 2. *Operator* τ *is defined also for sets. If* $A \subseteq C_{iw}$,

$$\tau(A) = \{\ \tau(S)\ :\ S \in A\ \}. \tag{3}$$

Definition 3. *If* $S_1, S_2 \in C_{iw}$, *then* $S_1 \leq S_2 \Longleftrightarrow S_1 \in \tau(S_2)$.

Lemma 1. *Definition 3. is a partial ordering on* C_{iw}, $\langle C_{iw}, \leq \rangle$ *is a poset.*

The intention is to construct lattices to elements of C_{iw}, which provide versatile tools to characterize the underlying words. If considering a word S^n, the base set of the sought lattice is generated by $\tau(\tau(S^n))$. This is the set of sub-word sets assigned individually to elements of the sub-word set $\tau(S^n)$. However the base set of lattice algebra $\langle\overline{\tau(\tau(S^n))},\wedge,\vee\rangle$ is the closure set $\overline{\tau(\tau(S^n))}$ obtained by applying the lattice operations onto $\tau(\tau(S^n))$ (for notational conventions see Appendix 2.). The lattice operations are defined in general terms for elements $a,b,c,...\in\tau(\tau(S^n))$, which means that elements of the base set are sets.

Definition 4. *The join* $(\vee)$ *and meet* $(\wedge)$ *will be,*

$$a\vee b=a\cup b,\quad a\wedge b=a\cap b. \tag{4}$$

The result of the join is the union of operands, that of the meet is the intersection. A finite application of the above operations yields the closure set $\Omega\equiv\overline{\tau(\tau(S^n))}$. To be short in the sequel, elements of Ω will be denoted by the letters $x,y,z,...$ and the algebra as word-lattice $\mathbf{L}(\Omega)=\langle\overline{\tau(\tau(S^n))},\wedge,\vee\rangle$. The least $(\hat{0}^{\Omega})$ and greatest $(\hat{1}^{\Omega})$ elements of $\mathbf{L}(\Omega)$ are $\hat{0}^{\Omega}=\bigwedge x$ and $\hat{1}^{\Omega}=\bigvee x$. The elements just above $\hat{0}^{\Omega}$ are the atoms of the lattice associated with letters of the alphabet. Sometimes for illustration the generating word will be written directly, as $x\equiv<AGGCT>,\ y\equiv<TACCU>,...$, where $<AGGCT>\equiv\tau(\psi(AGGCT))$, etc.. By Definitions 3. and 4., for those elements of Ω, which are of form $x=\tau(S_1),\ y=\tau(S_2)\in\Omega$,

$$x\wedge y=x\Longrightarrow S_1\leq S_2,\quad x\vee y=y\Longrightarrow S_1\leq S_2. \tag{5}$$

Therefore, elements of Ω can be written as follows $\tau(<w_1>)\cup...\cup\tau(<w_r>)$, where $w_i\in C_{ch},\ 1\leq i\leq r$ and for all $k,l,\ k\neq l,\ w_k\not\subseteq w_l$ and $w_l\not\subseteq w_k$.

Lemma 2. *Word-lattice* $\mathbf{L}(\Omega)$ *is distributive.*

Unfortunately the Hasse-diagram of $\mathbf{L}(\Omega)$ is generally too complicated to be drawn, therefore the $S^4(14,17)=\psi(AUAU)$ sub-word of (1) is separated together with the homogenous sub-word $S^4(19,22)=\psi(UUUU)$ and their associated Hasse-diagrams are displayed on figure 1.. Taking into account $AUAU$, elements just above $\hat{0}$ are the atoms $<A>,\ <U>$ of the lattice. The covering elements are images of disconnected sub-sequences $<A>\vee<U>$. The next level contains images of words $<AU>$ and $<UA>$ and in the next higher level again non-words follow. The dual atom element is generally a non-word, while the unit element refers to the $<AUAU>$ image of the full word. If the nucleotide chain is built homogenously, such as $UUUU$, $\mathbf{L}(\Omega)$ is a chain lattice. The length of longest maximal chain of the lattice depends

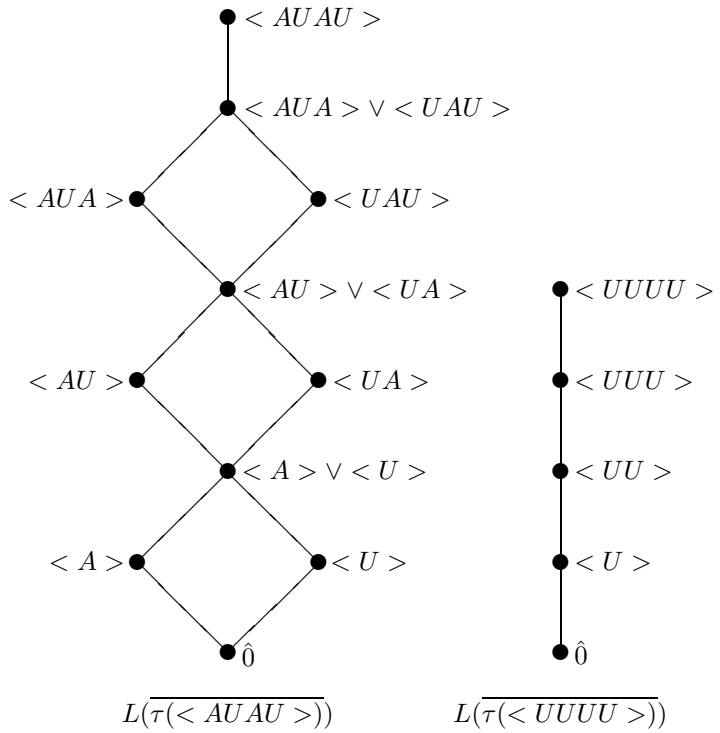


FIGURE 1. Hasse-diagrams of $\mathbf{L}(\overline{\tau(< AUAU >)})$ and $\mathbf{L}(\overline{\tau(< UUUU >)})$.

on the length and diversity of the full word and the length of longest anti-chain depends on its diversity. By the definition of lattice operations and that of the words, clearly not all elements of the lattice are assigned to words.

Lemma 3. *An element of* $\mathbf{L}(\Omega)$ *is join-irreducible if and only if it is a word.*

Theorem 1. *The partially ordered set of join-irreducible elements in* $\mathbf{L}(\Omega)$ *uniquely determines the word-lattice, therefore* $\mathbf{L}(\Omega_1) \cong \mathbf{L}(\Omega_2) \Longleftrightarrow \langle \tau(S_1^n), \leq \rangle \cong \langle \tau(S_2^n), \leq \rangle$.

Along any ascending chain of $\mathbf{L}(\Omega)$, join-irreducible elements represent the monotone 'construction' of the nucleotide chain, which is born from a single letter. The various chains of the lattice belong to differently positioned nucleotides as germinating centers.
The investigation of congruencies of an algebra generally enlightens its inherent structure.

Definition 5. *A congruency denoted* λ*, which is generated by the lattice interval* $[v : w]$ *means for elements* $x, y, v, w \in \mathbf{L}(\Omega)$*,*

$$x \equiv y \ (mod\ \lambda) \iff x \wedge v = y \wedge v, \quad x \vee w = y \vee w, \tag{6}$$

where the lattice interval $[v : w]$ *is the set of elements* $\{\ x\ :\ v \leq x \leq w\ \}$.

Lemma 4. *Definition 5. yields a congruence relation.*

The lattice interval is generated by the intersection of the 'principal ideal' $(w]$ and 'principal filter' $[v)$ (see Appendix 2.). A congruence relation will be denoted shortly as λ (or λ_q) and the congruent elements as $x \equiv y \ (mod\ \lambda)$. We call 'basis' congruence relations those, which are generated by a single lattice interval $\lambda \sim [v : w]$, where $\sim$ means 'associated with'. Elements congruent with a given $x \in \Omega$ are in the congruence class x/λ.

Lemma 5. *The congruence classes form convex sublattices in* $\mathbf{L}(\Omega)$.

The congruence relations themselves can be ordered by inclusion obtaining the lattice of congruencies $\mathbf{Con}(\mathbf{L}(\Omega))$ (see Appendix 2.). The operations $\lambda_i \wedge \lambda_j$ and $\lambda_i \vee \lambda_j$ are defined in the usual ways [24].

Lemma 6. *Lattice* $\mathbf{Con}(\mathbf{L}(\Omega))$ *is distributive.*

A given congruence relation partitions a lattice collapsing several elements into disjoint classes and yielding the factor lattice $\mathbf{L}(\Omega)/\lambda$. The zero element $(\hat{0}^C)$ collapses zero-length intervals inducing single-element congruence classes, the atoms collapse two elements, while the unit element $(\hat{1}^C)$ of $\mathbf{Con}(\mathbf{L}(\Omega))$ collapses all elements into a single class,

$$if\ x = y\ (x \equiv y\ (mod\ \hat{0}))\ and\ \forall x, y\ (x \equiv y\ (mod\ \hat{1})). \tag{7}$$

For the structure of the word, the composition of congruence classes bears importance. Congruence class x/λ includes x and all y having mutual meet and join in the lattice interval determined by λ. The join-irreducible elements (images of words) of the lattice and the reducible elements are not of equivalent value for the construction of nucleotide chain. Therefore the concept of the 'skeleton class' (SC) is introduced,

$$\hat{x}/\lambda \equiv SC(x/\lambda) := \{\ S_0 \in \tau(S^n)\ :\ \exists y((y \equiv x\ (mod\ \lambda)) \bigwedge (y = \tau(S_0))\}. \tag{8}$$

This set contains only the pre-images of join-irreducible elements in x/λ. Every ascending chain in $\mathbf{Con}(\mathbf{L}(\Omega))$ generates a sequence of SC-s by the following lemma.

Lemma 7. $SC(x/\lambda_q) \subseteq SC(x/\lambda_p)\ if\ \lambda_q \leq \lambda_p$.

The congruence classes and SC-s assemble those elements, which are inherently related by the lattice structure. The relationship between two words S_1 and S_2 is characterized qualitatively by the mutual mapping properties of associated word-lattices $\mathbf{L}(\Omega_1)$ and $\mathbf{L}(\Omega_2)$. In this report our interest is focused onto different kinds of isomorphisms, while the weaker mapping properties are discussed separately [25].

3. Isomorphism with condition

When comparing two words by the mapping properties of their associated word algebras, the expectable information refers to the similarity of the internal structures of the words, but not to their direct correspondence. The associated word algebras may have isomorphic, homomorphic maps into each other or partial embeddings, but these concepts do not imply the matching of the words and the obtained information is incomparable in nature with the usual matching percentages.

Definition 6. *Two words are isomorphic, if their associated word-lattices are isomorphic.*

Definition 7. *Two lattices are isomorphic* $\mathbf{L}(\Omega_1) \cong_\phi \mathbf{L}(\Omega_2)$ *with bijection* ϕ*, if*

$$\phi(x \wedge y) = \phi(x) \wedge \phi(y), \quad \phi(x \vee y) = \phi(x) \vee \phi(y), \quad x, y \in \Omega_1. \tag{9}$$

There is a variety of isomorphic words, but the relation between isomorphism and the structures of underlying words is not transparent, but rather deep-lying.

Definition 8. *Transposition operation* π_{ab} *interchanges elements* a, b *in the integer alphabet,*

$$\pi_{a,b}(s_1 ... s_n) = l_1 ... l_n, \quad l_i = \begin{cases} b, & if\ s_i = a, \\ a, & if\ s_i = b, \\ s_i, & otherwise. \end{cases} \tag{10}$$

Reflection operation ρ *mirrors the indexes of elements* $\rho(s_1 ... s_n) = s_n ... s_1$.

Lemma 8. *Transposing two letters in the integer alphabet generates an isomorphic word.*

Since the transposition of two letters does not change the structure of word-lattice, words containing all four elements of the alphabet have 4! isomorphic counterparts from the possible permutations of nucleotides. The complementation of a double-stranded DNA-chain is one realization of these transformations.

Lemma 9. *Reflection of a word generates an isomorphic word.*

Lemma 10. *The reflection and transposition operations form the transformation group* $\langle G, \cdot \rangle$*, where* $G = \{\pi_{a,b} \ : \ a, b \in \textit{'alphabet'} \ \} \cup \{\rho\}$

Although under the action of transformation group $\langle G, \cdot \rangle$ isomorphic words are created, the operations allow only a tight relationship of the words. The considered operations are global in the sense that their effect is spread onto the whole word. In spite of this the genetical mechanism of mutation is rather local. Therefore the unrestricted isomorphism is too strong a relationship for nucleotide chains and we introduce two weaker concepts.

Distributive lattices have several useful properties, for instance a height function is defined. If x is an element of the distributive lattice, $h(x)$ is the length of longest maximal chain in the lattice interval $[\hat{0}, x]$, (while another statement ensures to be every maximal chain of the same length [24].)

Definition 9. *The lower partial isomorphism* ($p_{\underline{r}} - isomorphism$) $(\mathbf{L}(\Omega_1) \cong_{\phi,\underline{r}} \mathbf{L}(\Omega_2))$ *of degree* r *and bijection* ϕ *requires the isomorphism of partial lattices [24]* $\mathbf{L}_p(H_{1,\underline{r}}) \cong_\phi \mathbf{L}_p(H_{2,\underline{r}})$, *where the base sets of lattices are restricted to subsets of* Ω_1 *and* Ω_2, *respectively with elements* $H_{1,\underline{r}} := \{x : h(x) \leq r\} \subseteq \Omega_1,\ H_{2,\underline{r}} := \{y : h(y) \leq r\} \subseteq \Omega_2$,

$$\phi(x \wedge y) = \phi(x) \wedge \phi(y), \quad \phi(x \vee y) = \phi(x) \vee \phi(y), \quad \forall x, y, x \vee y \in H_{1,\underline{r}}. \tag{11}$$

The upper partial isomorphism ($p_{\overline{s}}-isomorphism$) $(\mathbf{L}(\Omega_1) \cong_{\phi,\overline{s}} \mathbf{L}(\Omega_2))$ *of degree* s *and bijection* ϕ *sets similar requirements,* $H_{1,\overline{s}} := \{x : h(x) \geq h(\hat{1}_1) - s\} \subseteq \Omega_1,\ H_{2,\overline{s}} := \{y : h(y) \geq h(\hat{1}_2) - s\} \subseteq \Omega_2$,

$$\phi(x \wedge y) = \phi(x) \wedge \phi(y), \quad \phi(x \vee y) = \phi(x) \vee \phi(y), \quad \forall x, y, x \wedge y \in H_{1,\overline{s}}. \tag{12}$$

By a little lose, qualitative explanation $p_{\underline{r}}$-isomorphism and $p_{\overline{s}}$-isomorphism mean that subwords assigned to elements $h(x), h(x') \leq r$ or $h(x) \geq h(\hat{1}_1) - s$ and $h(x') \geq h(\hat{1}_2) - s$ and $(x \in \mathbf{L}(\Omega_1),\ x' \in \mathbf{L}(\Omega_2))$ are 'arranged' among themselves in the same ways in both lattices. The p-isomorphism allows local changes in the nucleotide chain providing a realistic approach to describe algebraic effects of biological mutation mechanism.

Theorem 2. *If* $\mathbf{L}(\Omega_1) \cong_{\phi_1,\underline{r}} \mathbf{L}(\Omega_2)$ *and* $\mathbf{L}(\Omega_1) \cong_{\phi_2,\overline{s}} \mathbf{L}(\Omega_2)$ $(h(\hat{1}_1) = h(\hat{1}_2) = r + s)$, *then the simultaneous* $p_{\underline{r}}$*-isomorphism and* $p_{\overline{s}}$*-isomorphism imply the unrestricted isomorphism of lattices, if the mappings* $\phi_1,\ \phi_2$ *are identical for the common elements,*

$$\Big(\mathbf{L}(\Omega_1) \cong_{\underline{r}} \mathbf{L}(\Omega_2)\Big) \bigwedge \Big(\mathbf{L}(\Omega_1) \cong_{\overline{s}} \mathbf{L}(\Omega_2)\Big) \bigwedge \Big(\forall x(\phi_1(x) = \phi_2(x),\ if\ h(x) = r)\Big) \iff \mathbf{L}(\Omega_1) \cong \mathbf{L}(\Omega_2). \tag{13}$$

If given $\mathbf{L}(\Omega_1)$ and $\mathbf{L}(\Omega_2)$ the maximal degrees of partial isomorphisms $r_{max} := \max_r\{r : \mathbf{L}(\Omega_1) \cong_{\underline{r}} \mathbf{L}(\Omega_2)\}$ and $s_{max} := \max_s\{s : \mathbf{L}(\Omega_1) \cong_{\overline{s}} \mathbf{L}(\Omega_2)\}$ can be determined. The differences $\Delta_1 = (h(\hat{1}_1) - s) - r$ and $\Delta_2 = (h(\hat{1}_2) - s) - r$ are closely related to the tightness of relationship of the underlying nucleotide chains. While unrestricted isomorphism ($\Delta_1 = \Delta_2 = 0$) (under the conditions of Theorem 2.) allows a relatively small number of isomorphic, but different nucleotide chains ($2 \cdot 4!$), $\Delta > 0$ difference implies a rapidly growing set of partially isomorphic

and different chain molecules. However unrestricted isomorphism is a strong requirement for the structures of underlying chains, still does not ensure complete matching and p-isomorphism allows much more freedom. Therefore a measure is needed to define the 'similarity distance' of two words and the next section is devoted to this question.

4. Shape lattice and measure

While nucleotide chains are compared primarily by the mapping properties of associated lattices $\mathbf{L}(\Omega_i)$ and $\mathbf{Con}(\mathbf{L}(\Omega_i))$, beyond the comparison of associated algebras practical demands also require a direct comparison of two words. This direct comparison should be detailed and well structured, which means that an individual 'measure of correspondence' is needed for every relevant set of sub-words, instead of a global parameter for the whole word.

In any distributive lattice, one can define a non-negative function μ with the following properties. If x, y are elements of the lattice, then

$$x = y \Longrightarrow \mu(x) = \mu(y), \tag{14}$$

$$x \wedge y = \hat{0} \Longrightarrow \mu(x \vee y) = \mu(x) + \mu(y). \tag{15}$$

Under these circumstances μ is a finitely additive measure and it is strictly positive, if

$$\mu(x) = 0 \Longrightarrow x = \hat{0}. \tag{16}$$

Though $\mathbf{L}(\Omega)$ is distributive, if rejecting non-words and retaining only words in the congruence classes (SC), the remaining part of the base set can not be ordered into a distributive lattice. To remedy this difficulty, the sceleton classes are mapped onto a set of n-vectors. The closure of this vector set is lattice-ordered into a distributive lattice, which is finally furnished with a measure.

Definition 10. *For every* $\hat{x}/\lambda_q,\ x \in \mathbf{L}(\Omega),\ \lambda \in \mathbf{Con}(\mathbf{L}(\Omega))$,

$$\alpha_{\hat{x}/\lambda_q}(i) = \begin{cases} 1, & \text{if exists indexes } k,l,\ 1 \le k \le i \le l \le n \text{ and } S^n(k,l) \in \hat{x}/\lambda_q, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

and the n*-vector* $\vec{\alpha}_{\hat{x}/\lambda_q} \equiv [\alpha_{\hat{x}/\lambda_q}(1), ..., \alpha_{\hat{x}/\lambda_q}(n)]^T$ *is called 'shape characteristic vector'.*

Definition 11. *For every* $\hat{x}/\lambda_q,\ x \in \mathbf{L}(\Omega),\ \lambda \in \mathbf{Con}(\mathbf{L}(\Omega))$,

$$\beta_{\hat{x}/\lambda_q}(i) = \begin{cases} s_i \equiv S^n(i,i), & \text{if exists indexes } k,l,\ 1 \le k \le i \le l \le n \text{ and } S^n(k,l) \in \hat{x}/\lambda_q, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

and the n*-vector* $\vec{\beta}_{\hat{x}/\lambda_q} \equiv [\beta_{\hat{x}/\lambda_q}(1), ..., \beta_{\hat{x}/\lambda_q}(n)]^T$ *is called 'shape property vector'.*

Vector $\vec{\alpha}$ contains 1-s in all those positions, where letters of the words of the given SC are located, while $\vec{\beta}$ contains directly the letters. The images of all congruence classes are the vector sets $E = \{\vec{\alpha}_{\hat{x}/\lambda_q}\}$ and $F = \{\vec{\beta}_{\hat{x}/\lambda_q}\}$. Lattices are constructed over E and F, by the following operations. (To proceed in a formal generality, the operands will be denoted as $\vec{a}, \vec{b}, \vec{c}, ...$)

Definition 12. *The meet and join are constructed component-by-component, as the minimum and the maximum of operands,*

$$\vec{a} \wedge \vec{b} = \vec{c},\ c_i := min\{a_i, b_i\}, \quad \vec{a} \vee \vec{b} = \vec{c},\ c_i := max\{a_i, b_i\},\ i \in I. \tag{19}$$

The order relations and lattice operations are connected in the usual way (see Appendix 2.). The closure sets $\overline{E},\ \overline{F}$ are obtained by finite applications of the specified operations. The constructed lattices, $\mathbf{L}(\overline{E}) = \langle \overline{E}, \wedge, \vee \rangle$ and $\mathbf{L}(\overline{F}) = \langle \overline{F}, \wedge, \vee \rangle$ are characteristic to the generalized 'shape' of the underlying nucleotide chain and will be called 'shape characteristic lattice' and 'shape property lattice'. The universal bounds arise by the operations $\hat{0}^{\overline{E}} = \bigwedge_{\vec{a} \in \overline{E}} \vec{a}$ and $\hat{1}^{\overline{E}} = \bigvee_{\vec{a} \in \overline{E}} \vec{a}$, as similarly $\hat{0}^{\overline{F}} = \bigwedge_{\vec{b} \in \overline{F}} \vec{b}$ and $\hat{1}^{\overline{F}} = \bigvee_{\vec{b} \in \overline{F}} \vec{b}$, $\vec{a} \in \overline{E},\ \vec{b} \in \overline{F}$.

Lemma 11. *The shape characteristic and shape property lattices are distributive.*

Theorem 3. *The isomorphism of word-lattices implies the isomorphism of shape characteristic and shape property lattices and the reverse implication is true as well,*

$$(\mathbf{L}(\Omega_1) \cong \mathbf{L}(\Omega_2)) \Longleftrightarrow ((\mathbf{L}(\overline{E}_1) \cong \mathbf{L}(\overline{E}_2)) \bigwedge (\mathbf{L}(\overline{F}_1) \cong \mathbf{L}(\overline{F}_2))). \tag{20}$$

Lemma 12. *Function*

$$\mu(\vec{\alpha}_{\hat{x}/\lambda_q}) := \vec{\alpha}_{\hat{x}/\lambda_q} \circ \vec{\alpha}_{\hat{x}/\lambda_q}, \quad \vec{\alpha}_{\hat{x}/\lambda_q} \in \overline{E},\ \lambda_q \in \mathbf{Con}(\mathbf{L}(\Omega_1)). \tag{21}$$

(*where* $\circ$ *denotes the scalar product*) *is a strictly positive measure on the distributive lattice* $\mathbf{L}(\overline{E})$.

If normalizing the vectors $(\vec{\alpha}_{\hat{x}/\lambda_q}/||\vec{\alpha}_{\hat{x}/\hat{1}^{\overline{E}}}||)$, measure $\mu(\vec{\alpha}_{\hat{x}/\lambda_q}/||\vec{\alpha}_{\hat{x}/\hat{1}^{\overline{E}}}||)$ becomes a density function. It provides information on the distribution of considered property. For example, the density function value $\mu((\vec{\alpha}_{\hat{x}/\hat{0}^{\overline{E}}} \bigwedge \vec{\alpha}_{\hat{<}A>/\hat{0}^{\overline{E}}})/||\vec{\alpha}_{\hat{x}/\hat{1}^{\overline{E}}}||)$ is the relative A-content of the word $b \in \tau(S),\ x = \tau(b)$.

The correspondence of two differing nucleotide chains of the same lengths is characterized by their 'similarity distance'. This measure is based on the number of differing nucleotides. Let be given two n-vectors ${}^1\vec{\beta}_{\hat{x}/\lambda_p} \in \overline{F}_1$ and ${}^2\vec{\beta}_{\hat{y}/\lambda_q} \in \overline{F}_2$ and the following 'characteristic operator' κ,

Definition 13.

$$\kappa(\vec{a}) = \vec{b}, \quad \vec{b} = \begin{cases} b_i = 1, & if\ a_i \neq 0, \\ 0, & otherwise. \end{cases} \tag{22}$$

where $\vec{a},\ \vec{b}$ *are two* n*-vectors in general notations.*

The characteristic operator assigns 1 to every non-zero element of the vector, therefore the difference vector $({}^1\vec{\beta}_{\hat{x}/\lambda_p} - {}^2\vec{\beta}_{\hat{y}/\lambda_q})$ will contain 1-s in every position, where the component vectors do not match and 0-s, where do. The measure of 'similarity distance' depends on the number of non-matching elements.

Lemma 13. *Function*

$$\mu(\kappa({}^1\vec{\beta}_{\hat{x}/\lambda_p} - {}^2\vec{\beta}_{\hat{y}/\lambda_q})) := \kappa({}^1\vec{\beta}_{\hat{x}/\lambda_p} - {}^2\vec{\beta}_{\hat{y}/\lambda_q}) \circ \kappa({}^1\vec{\beta}_{\hat{x}/\lambda_p} - {}^2\vec{\beta}_{\hat{y}/\lambda_q}) \tag{23}$$

$${}^1\vec{\beta}_{\hat{x}/\lambda_p} \in \overline{F}_1,\ {}^2\vec{\beta}_{\hat{y}/\lambda_q} \in \overline{F}_2,\ \lambda_p \in \mathbf{Con}(\mathbf{L}(\Omega_1)),\ \lambda_q \in \mathbf{Con}(\mathbf{L}(\Omega_2)).$$

satisfies the conditions to be a metric.

The constructed similarity distance is a 'measure of correspondence' for every individual pair of SC-s. All similarity distances, which are assigned to a fixed (x, y) couple of elements, but systematically to every congruence relation form the 'Similarity Distance' matrix

$$SD(x, y) := [\mu(\kappa({}^1\vec{\beta}_{\hat{x}/\lambda_p} - {}^2\vec{\beta}_{\hat{y}/\lambda_q}))]_{p,q},\ x \in \mathbf{L}(\Omega_1),\ y \in \mathbf{L}(\Omega_2). \tag{24}$$

If the associated lattices are isomorphic ($\mathbf{L}(\Omega_1) \cong_\phi \mathbf{L}(\Omega_2)$), the diagonal elements (for which $p = q$) of the matrix $SD(x, \phi(x)),\ x \in \mathbf{L}(\Omega_1),\ \phi(x) \in \mathbf{L}(\Omega_2)$ measure the similarity distances of the isomorphism coupled SC-s. That element in each matrix $SD(x, y)$ for which $\lambda_p = \hat{1}_1^{\overline{F}}$ ($\lambda_q = \hat{1}_2^{\overline{F}}$ respectively) agrees for every $x \in \mathbf{L}(\Omega_1)$ and $y \in \mathbf{L}(\Omega_2)$ providing the '*g*lobal *s*imilarity *d*istance',

$$gsd_{F_1,F_2} := \mu_\beta({}^1\vec{\beta}_{\hat{x}/\hat{1}^{\overline{F}_1}} - {}^2\vec{\beta}_{\hat{y}/\hat{1}^{\overline{F}_2}}), \quad x \in \mathbf{L}(\overline{\Omega}_1),\ y \in \mathbf{L}(\overline{\Omega}_2) \tag{25}$$

characterizing the overall correspondence of two selected words. While matrix $SD(x, y)$ collects all similarity distances referring to the fixed lattice elements (x, y) and all congruence relations (λ_p, λ_q), the total available metric information is collected in the hyperlattice TSD 'Total Similarity Distance',

$$TSD := [SD(x, y)]_{x,y},\ x \in \mathbf{L}(\overline{\Omega}_1),\ y \in \mathbf{L}(\overline{\Omega}_2). \tag{26}$$

This matrix comprises similarity distances between all congruence classes. Because β-measure depends on the choice of alphabet, only words over identical alphabets can be compared. Although these last considerations referred to vectors associated with different lattices, all considerations apply to vectors of the same lattice, too.

As to the meaning of obtained information, the established lattice algebra orders the images of words over a fixed alphabet. The obtained lattice $\mathbf{L}(\Omega)$ depends on the coupling mode and diversity of the words. For example, if taking into account a join-irreducible element of the lattice, the number of upper closest join-irreducibles depend on the number of different words including the selected one considering every occurrence of it along the whole chain. The obtained lattice is decomposed into sub-algebras, which are congruence classes composed of specifically relating elements of the lattice. The distribution of any lattice property expressible by lattice polynomials [24] is characterized by a density function, the mutual correspondence of congruence classes is measured by an appropriate similarity distance, which provides metric informations for the local, as well as global correspondences of the compared nucleotide chains.

5. Illustrations and conclusions

In this section some simple examples are given to elucidate advantages of lattice theoretical characterization of nucleotide chains over simple string matching schemes. In our opinion, even though direct matching of nucleotide chains is important since it defines identical transcripts, the relationship of lattice-based 'global structures' of two nucleotide chains say more then mere matching percentage.

Since word-lattices are generally too large for displaying the Hasse-diagrams, we restrict illustrations to lattices associated with small portions of the following nucleotide sequences,

$$C_a.)\ TTAGGGTTAGGGTTAGGG, \quad C_b.)\ AATCCCAATCCCAATCCC,$$

$$C_c.)\ TTCCATTCCATTCCATTCCA.$$

We do not attempt to present 'results of practical utility', because the necessary huge lattices would totally destroy our goal to make us understood. Therefore the forthcoming examples serve exclusively purposes of illustration of some selected ideas of this report.

The first two words C_a and C_b are complementary, the largest repeating elementary units consist of six letters. Their word-lattices are isomorphic by Lemma 8., but figure 2. displays only the Hasse-diagrams of corresponding 4-letter units $AGGG$ and $TCCC$. Choosing the sceleton class $SC(x/\lambda)$, $x =< AG >$, $\lambda \sim [< A > \vee < G > : < AG > \vee < GG >]$ in $\mathbf{L}(\overline{\tau(< AGGG >)})$, the associated vectors are $\vec{\alpha}_{\hat{x}/\lambda} = [1, 1, 0, 0]^T$ and $\vec{\beta}_{\hat{x}/\lambda} = [1, 2, 0, 0]^T$. In $\mathbf{L}(\overline{\tau(< TCCC >)})$ the

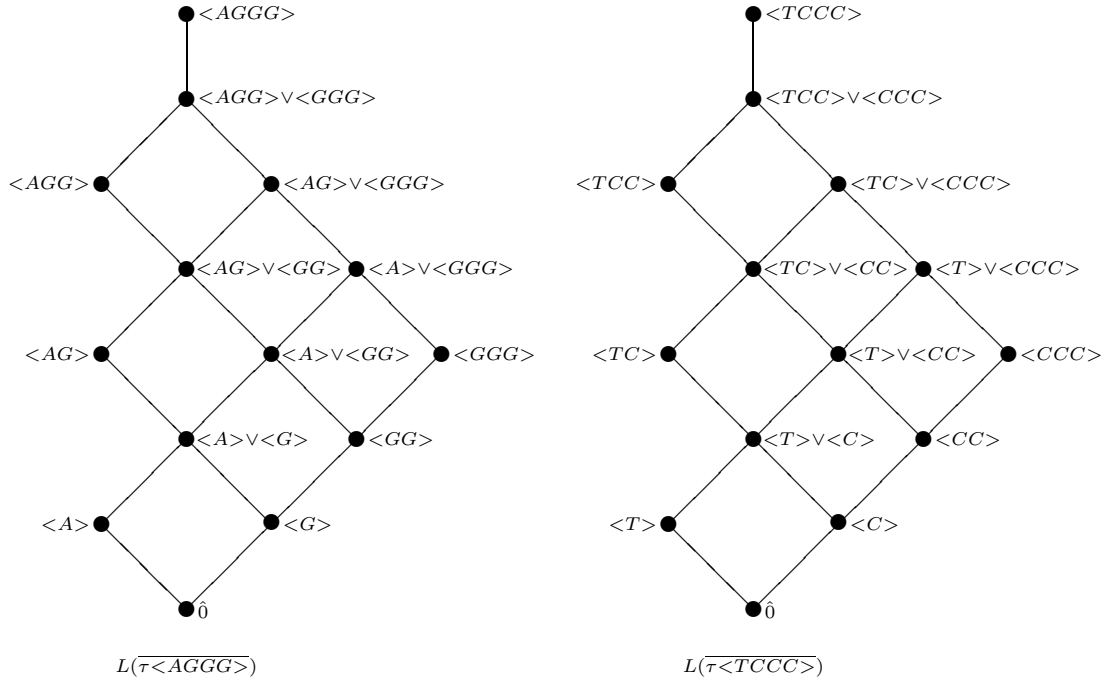


FIGURE 2. Complementary words with isomorphic word-lattices.

corresponding sceleton class and vectors are $SC(x'/\lambda')$, $x' =< TC >$, $\lambda' \sim [< T > \vee < C > : < TC > \vee < CC >]$ and $\vec{\alpha}_{\hat{x}'/\lambda'} = [1,1,0,0]^T$, $\vec{\beta}_{\hat{x}'/\lambda'} = [4,3,0,0]^T$. The number of differing letters is provided by β-measure $\mu(\kappa({}^1\vec{\beta}_{\hat{x}/\lambda} - {}^2\vec{\beta}_{\hat{x}'/\lambda'})) = [1,1,0,0] \circ [1,1,0,0] = 2$. Though direct correspondence does not exist between the words, their internal structures are of the same construction. The images of various repeat units in both words form sub-lattices of $\mathbf{L}(\overline{\tau(\tau(\psi(C_a)))})$ and $\mathbf{L}(\overline{\tau(\tau(\psi(C_b)))})$ and the isomorphism forces pairing of corresponding units.

The $CCCA$ unit of C_b may result from $AGGG$ of C_a by a reflection $\rho(AGGG) = GGGA$ and a subsequent transposition $\pi_{G,C}(GGGA) = CCCA$. (For the sake of simple notation the operators were applied directly to the letters). The Hasse-diagrams corresponding to the words $AGGG$ and $GGGA$ are shown on figure 3.

If taking into account C_c also, the isomorphism of full words is clearly lost, because even their lengths are different. However there are several isomorphic sub-lattices and corresponding sub-words. The largest elementary repeat unit $CCATT$ of C_c is isomorphic with $TTAGG$ of C_a. It is generated by transpositions $\pi_{C,G}(\pi_{C,T}(CCATT)) = TTAGG$. $CCATT$ of S_c is also isomorphic with $AATCC$ of S_b, when it is generated by the transpositions $\pi_{C,A}(\pi_{A,T}(CCATT)) = AATCC$.

Finally the partial isomorphism is illustrated by the examples $AGAG$, $AGAGAG$ and $AGGG$, $AAGG$, where most of them have isomorphic counterparts in C_a, C_b or C_c. The

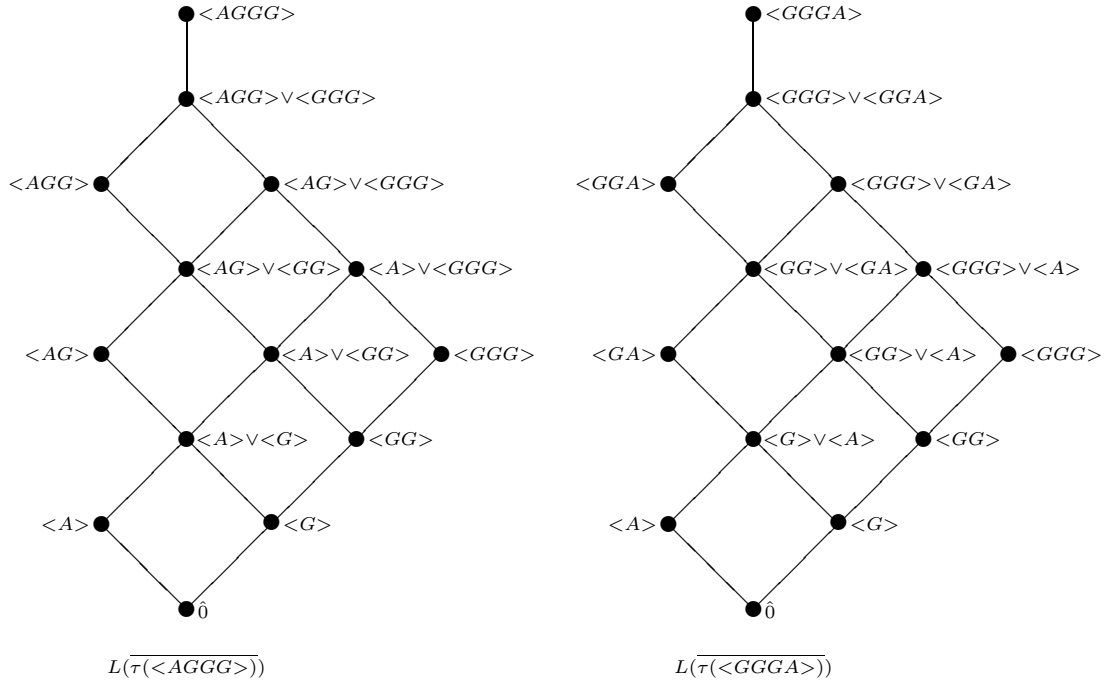


FIGURE 3. Mirror image words with isomorphic word-lattices.

first two words contain the same 2-letter elementary repeat unit and differ only in their full lengths. Their lattices are displayed on figure 4.. It is clear that up to $h(x) = 6$ height the two words have the same structures implying lower partially isomorphic lattices $\mathbf{L}(\Omega_1) \cong_{\underline{6}} \mathbf{L}(\Omega_2)$. The upper parts of the lattices are isomorphic as well implying the upper partial isomorphism $\mathbf{L}(\Omega_1) \cong_{\overline{7}} \mathbf{L}(\Omega_2)$. This means for example that sub-words $AGAGA$ and $GAGAG$ are in similar relation to each other as the sub-words AGA and GAG. Let us consider the word $AGAG$ and an ascending chain of congruence relations by selecting the generator lattice intervals to be of the following forms $\lambda_i \sim [\hat{0} : x_i]$ with the convention $i < j \Longrightarrow h(x_i) < h(x_j)$. The selected chain of congruencies is $\lambda_1 \sim [< \hat{0} >:< A > \vee < G >]$, $\lambda_2 \sim [< \hat{0} >:< AG > \vee < GA >]$, $\lambda_3 \sim [< \hat{0} >:< AGA > \vee < GAG >]$, $\lambda_4 \sim [< \hat{0} >:< \hat{1} >]$. The associated vectors are $\alpha_{<\hat{AG}>/\lambda_q} = [1,1,1,1]^T (q = 1, ..., 4)$ expressing the homogenous construction of the word. The density is therefore uniform with value $\mu(\vec{\alpha}_{<\hat{AG}>/\lambda_q}/||\vec{\alpha}_{<\hat{AG}>/\hat{1}^{\overline{E}}}||) = 1$. A similarly uniform distribution can be obtained for the word $AUAUAU$, too.

The next example refers to words $AGGG$ and $AAGG$. Their lattices are displayed on figure 5.. The lattice associated with the first word is already well known, the second looks quite complicated although only one position ($G \longrightarrow A$) has been changed. It is apparent that increasing the symmetry of the word, the symmetry of the lattice is also increased. The partial isomorphisms are significantly weaker then in the previous example and the lower partial

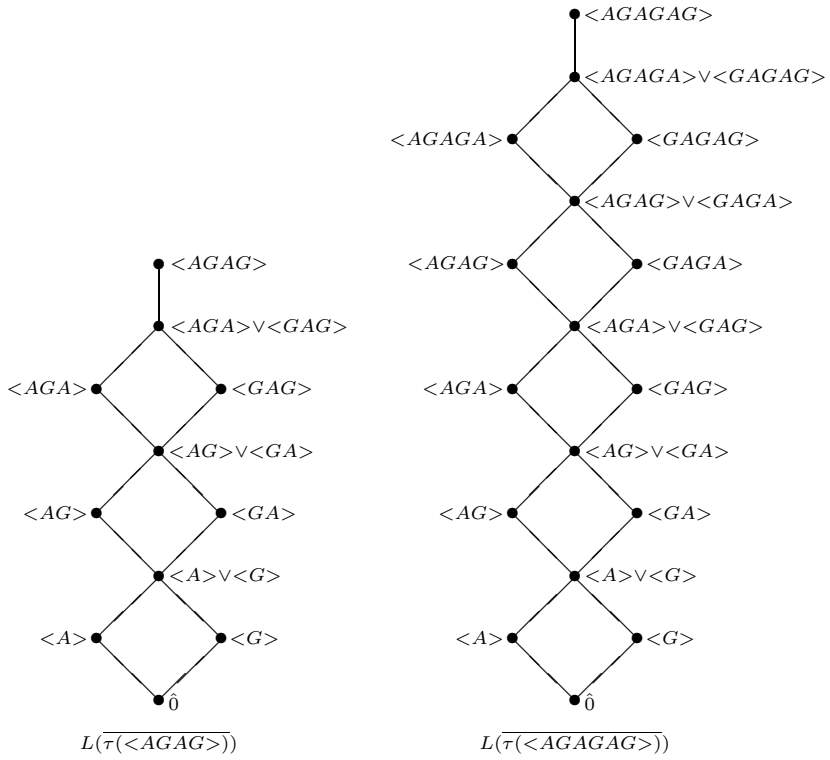
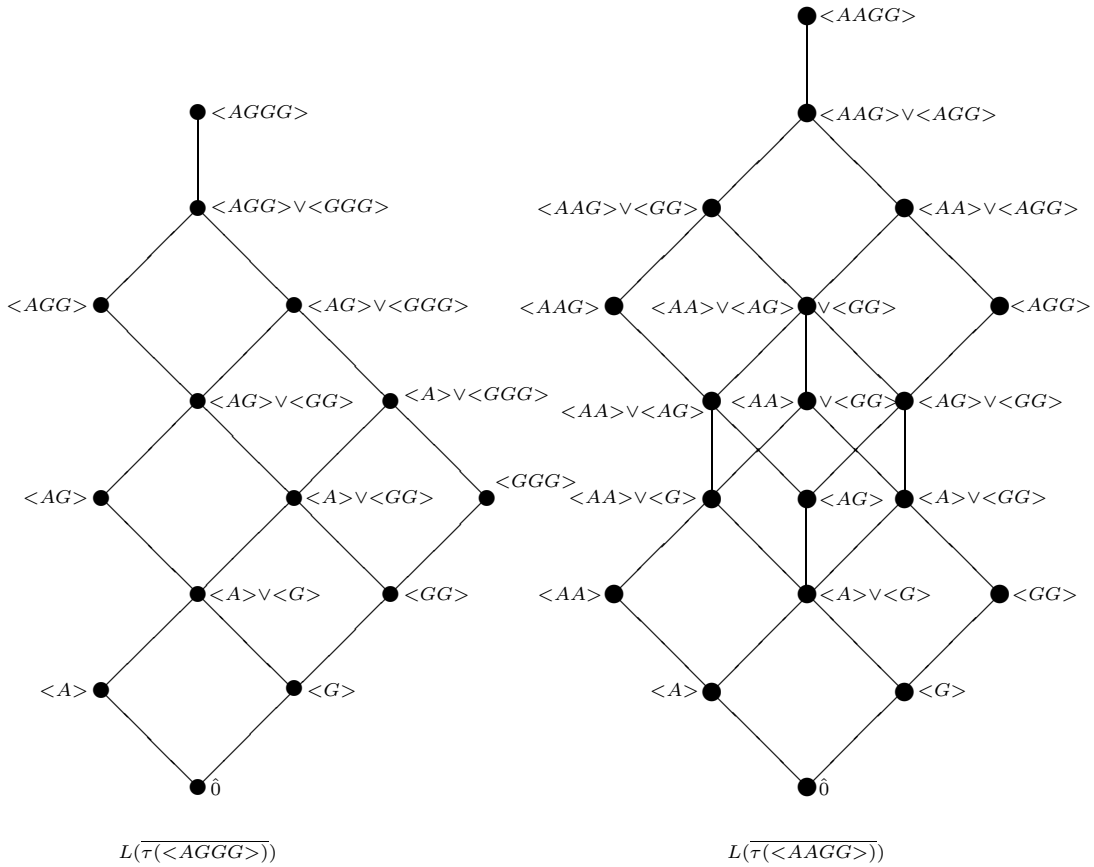


FIGURE 4. Lower and upper partial isomorphisms of word-lattices.

isomorphism is only of first degree, $\mathbf{L}(\Omega_1) \cong_{\underline{1}} \mathbf{L}(\Omega_2)$, while the upper partial isomorphism is of second degree $\mathbf{L}(\Omega_1) \cong_{\overline{2}} \mathbf{L}(\Omega_2)$. Beyond the partial isomorphism there are a number of isomorphic sub-lattices, as the intervals $[< G >:< AGG >]_1$ and $[< A >, < AAG >]_2$ or $[< G >, < AGG >]_1$ and $[< G >, < AGG >]_2$, what we call 'local isomorphism'. When investigating the distribution of the word AG, again an ascending chain of bracketing lattice intervals is chosen. For $AGGG$ the obtained congruence relation generating lattice intervals and congruence classes are listed in table 1.. The corresponding $\vec{\alpha}$ vectors and density function values are listed in table 2. The congruence relations and lattice intervals assigned to the word $AAGG$ are listed in table 3., while the corresponding $\vec{\alpha}$ vectors and density function values are listed in table 4.. As seen on tables 2. and 4., the distributions of $< \hat{AG} > /\lambda_q$ among the selected sceleton classes show basic similarity and minor differences in the words $AGGG$ and $AAGG$. If taking into account the full set or only a well chosen subset of lattice elements and congruence relations, the associated density functions characterize the word quite well.

Of course the above examples were very simple, but sufficiently complex for demonstrating the power of lattice-algebraic characterization of nucleotide chains. The global isomorphism of lattices, the local isomorphism of sub-lattices or the partial isomorphism of lattices give a colourful description of the chain structure, which is essentially different in nature from simple



$L(\overline{\tau(<AGGG>)})$ $L(\overline{\tau(<AAGG>)})$

FIGURE 5. Lower and upper partial isomorphisms of word-lattices.

$h(x)$	$\lambda \sim [v:w]$	$SC(< AG > /\lambda)$
2	$\lambda_{20} \sim [< \hat{0} >:< A > \vee < G >]$	$\{< AG >\}$
	$\lambda_{21} \sim [< \hat{0} >:< GG >]$	$\{< AG >\}$
3	$\lambda_{30} \sim [< \hat{0} >:< AG >]$	$\{< A >, < G >, < AG >\}$
	$\lambda_{31} \sim [< \hat{0} >:< A > \vee < GG >]$	$\{< AG >\}$
	$\lambda_{32} \sim [< \hat{0} >:< GGG >]$	$\{< AG >\}$
4	$\lambda_{40} \sim [< \hat{0} >:< AG > \vee < GG >]$	$\{< A >, < G >, < AG >, < GG >\}$
	$\lambda_{41} \sim [< \hat{0} >:< A > \vee < GGG >]$	$\{< AG >\}$
5	$\lambda_{50} \sim [< \hat{0} >:< AGG >]$	$\{< A >, < G >, < AG >, < GG >,$ $< AGG >\}$
	$\lambda_{51} \sim [< \hat{0} >:< AG > \vee < GGG >]$	$\{< A >, < G >, < AG >, < GG >,$ $< GGG >\}$
6	$\lambda_{60} \sim [< \hat{0} >:< AGG > \vee < GGG >]$	$\{< A >, < G >, < AG >, < GG >,$ $< GGG >, < AGG >\}$
7	$\lambda_{\hat{1}} \sim [< \hat{0} >:< AGGG >]$	$\{< A >, < G >, < AG >, < GG >,$ $< GGG >, < AGG >, < AGGG >\}$

TABLE 1. Selected congruence relations and generating lattice intervals of $\mathbf{L}(\overline{\tau(< AGGG >)})$.

$h(x)$	$\vec{\alpha}_{<\hat{AG}>/\lambda_q}$	$\mu(\vec{\alpha}/\|\|\vec{\alpha}\|\|)$	$\mu_{average}$
2	$[1,1,0,0]^T$	$1/2$	$1/2$
	$[1,1,0,0]^T$	$1/2$	
3	$[1,1,1,1]^T$	1	$2/3$
	$[1,1,0,0]^T$	$1/2$	
	$[1,1,0,0]^T$	$1/2$	
4	$[1,1,1,1]^T$	1	$3/4$
	$[1,1,0,0]^T$	$1/2$	
5	$[1,1,1,1]^T$	1	1
	$[1,1,1,1]^T$	1	
6	$[1,1,1,1]^T$	1	1
7	$[1,1,1,1]^T$	1	1

TABLE 2. Selected density function values associated with $\mathbf{L}(\overline{\tau(< AGGG >)})$.

$h(x)$	$\lambda \sim [v:w]$	$SC(< AG > /\lambda)$
2	$\lambda_{20} \sim [< \hat{0} >:< AA >]$	$\{< AG >\}$
	$\lambda_{21} \sim [< \hat{0} >:< A > \vee < G >]$	$\{< AG >\}$
	$\lambda_{22} \sim [< \hat{0} >:< GG >]$	$\{< AG >\}$
3	$\lambda_{30} \sim [< \hat{0} >:< AA > \vee < G >]$	$\{< AG >\}$
	$\lambda_{31} \sim [< \hat{0} >:< AG >]$	$\{< A >, < G >, < AG >\}$
	$\lambda_{32} \sim [< \hat{0} >:< A > \vee < GG >]$	$\{< AG >\}$
4	$\lambda_{40} \sim [< \hat{0} >:< AA > \vee < AG >]$	$\{< A >, < G >, < AA >, < AG >\}$
	$\lambda_{41} \sim [< \hat{0} >:< AA > \vee < GG >]$	$\{< AG >\}$
	$\lambda_{42} \sim [< \hat{0} >:< AG > \vee < GG >]$	$\{< A >, < G >, < AG >, < GG >\}$
5	$\lambda_{50} \sim [< \hat{0} >:< AAG >]$	$\{< A >, < G >, < AA >, < AG >,$ $< AAG >\}$
	$\lambda_{51} \sim [< \hat{0} >:< AA > \vee < AG > \vee < GG >]$	$< A >, < G >, < AG >, < AA >,$ $< GG >\}$
	$\lambda_{52} \sim [< \hat{0} >:< AGG >]$	$\{< A >, < G >, < GG >, < AG >,$ $< AGG >\}$
6	$\lambda_{60} \sim [< \hat{0} >:< AAG > \vee < GG >]$	$\{< A >, < G >, < AA >, < GG >,$ $< AG >, < AAG >\}$
	$\lambda_{61} \sim [< \hat{0} >:< AA > \vee < AGG >$	$\{< A >, < G >, < AA >, < GG >,$ $< AG >, < AGG >\}$
7	$\lambda_{70} \sim [< \hat{0} >:< AAG > \vee < AGG >]$	$\{< A >, < G >, < AA >, < GG >,$ $< AG >, < AAG >, < AGG >\}$
8	$\lambda_{\hat{1}} \sim [< \hat{0} >:< AAGG >]$	$\{< A >, < G >, < AA >, < GG >,$ $< AG >, < AAG >, < AGG >, < AAGG >\}$

TABLE 3. Selected congruence relations and generating lattice intervals of $\mathbf{L}(\overline{\tau(< AAGG >)})$.

matching comparisons. The multitude of distribution functions of properties expressible by lattice-polynomials provide all-round picture about the chain molecule. Finally the β-measure yields a vaste amount of deeply structured informations in essence similar to string matching based results, but significantly more detailed.

$h(x)$	$\vec{\alpha}_{<\hat{AG}>/\lambda_q}$	$\mu(\vec{\alpha}/\|\|\vec{\alpha}\|\|)$	$\mu_{average}$
2	$[0,1,1,0]^T$	1/2	1/2
	$[0,1,1,0]^T$	1/2	
	$[0,1,1,0]^T$	1/2	
3	$[0,1,1,0]^T$	1/2	2/3
	$[1,1,1,1]^T$	1	
	$[0,1,1,0]^T$	1/2	
4	$[1,1,1,1]^T$	1	5/6
	$[0,1,1,0]^T$	1/2	
	$[1,1,1,1]^T$	1	
5	$[1,1,1,1]^T$	1	1
	$[1,1,1,1]^T$	1	
	$[1,1,1,1]^T$	1	
6	$[1,1,1,1]^T$	1	1
	$[1,1,1,1]^T$	1	
7	$[1,1,1,1]^T$	1	1
8	$[1,1,1,1]^T$	1	1

TABLE 4. Selected density function values associated with $\mathbf{L}(\overline{\tau(< AAGG >)})$.

6. Appendix 1. (Proofs of the assertions)

Lemma 1.

Proof. Reflexivity and transitivity fulfills trivially, anti-symmetry follows from the simultaneous relations $(S_1 \leq S_2) \bigwedge (S_2 \leq S_1) \Longrightarrow S_1 = S_2$. Definition 3. is an order-relation, $\langle C_{iw}, \leq \rangle$ is a poset. □

Lemma 2.

Proof. The base set $\overline{\tau(\tau(S^n))}$ of $\mathbf{L}(\Omega)$ is the closure completion of $\tau(\tau(S^n))$ with all unions and intersections. Since the lattice operations agree with the set-theoretical union and intersection, the set-theoretical identities

$$a \cup (b \cap c) = (a \cup b) \cap (a \cup c)$$

$$a \cap (b \cup c) = (a \cap b) \cup (a \cap c)$$

are satisfied for all $a, b, c, \in \overline{\tau(\tau(S^n))}$. □

Lemma 3.

Proof. The statement follows directly from the definition of the word. Since a word $w \in \tau(S)$ by definition is a connected ensemble of juxtaposed letters, $\tau(w)$ is the image of a word, if $\tau(w) \in \tau(\tau(S))$. Join-irreducibility means $x \vee y = z \Longrightarrow z = x \; or \; z = y$, which means that no two elements exist in the principal ideal $(z]$, which are different from z, but generate it. Since $z = \tau(w)$ can not be generated by two elements, it must be the image of a word. In the

reverse order, the $z = \tau(w) \in \tau(\tau(S))$ element (the image of a word) can not be generated from two smaller elements and must be join-irreducible because the lattice operations are the set-theoretical union and intersection. □

Theorem 1.

Proof. It is supposed that $\mathbf{L}(\Omega_1)(\equiv \mathbf{L}(\overline{\tau(\tau(S_1))})) \cong_\phi \mathbf{L}(\Omega_2)(\equiv \mathbf{L}(\overline{\tau(\tau(S_2))}))$. Bijection ϕ uniquely and order-preserving couples images of words exclusively with images of words. For the relationship of order-relations and lattice-operations this implies the isomorphism of posets $\langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(S_2), \leq \rangle$ consisting exlusively in the words of the base sets of lattices. In the reverse order the isomorphism of full lattices $\mathbf{L}(\Omega_1) \cong_\phi \mathbf{L}(\Omega_2)$ follows from the isomorphism of posets $\langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(S_2), \leq \rangle$. This is because their base sets Ω_1 and Ω_2 are generated by the same order-preserving procedures $\tau(S_1) \longrightarrow \tau(\tau(S_1))$ and $\tau(S_2) \longrightarrow \tau(\tau(S_2))$, as well as identical order-presrving closure formation by lattice operations given in Definition 4.. □

Lemma 4.

Proof. The proof is found in [24]. □

Lemma 5.

Proof. Textbook [22] is referred. □

Lemma 6.

Proof. Textbook [24] is referred. □

Lemma 7.

Proof. Using Definition 5.,

$$\lambda_q \leq \lambda_p \Longrightarrow x/\lambda_q \leq x/\lambda_p \Longrightarrow \hat{x}/\lambda_q \leq \hat{x}/\lambda_p \ (\equiv SC(x/\lambda_q) \leq SC(x/\lambda_p)).$$

□

Lemma 8.

Proof. By Theorem 1., it is enough to show that $\langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(\pi_{a,b}(S_1)), \leq \rangle$. By Definition 8. the transposition operation is invertible, therefore the mapping $\phi : \tau(S_1) \longleftrightarrow \tau(\pi_{a,b}(S_1))$ is a bijection, $\forall w(\phi(w) := \pi_{a,b}(w),\ w \in \tau(S_1))$. Since the inclusion relation is preserved by the transposition operation, the mapping is order-preserving and the isomorphism of posets and lattices is maintained. □

Lemma 9.

Proof. The proof is along the same line as in Lemma 8.. □

Lemma 10.

Proof. There is a neutral element $\pi_{a,a} \equiv \pi_{b,b} \equiv ...$, the associativity follows trivially and because the operations $\pi,\ \rho$ are invertible, an inverse exists for every element. □

Theorem 2.

Proof. By assumption $\mathbf{L}(\Omega_1) \cong_{\phi_1,\underline{r}} \mathbf{L}(\Omega_2)$ and $\mathbf{L}(\Omega_1) \cong_{\phi_2,\overline{s}} \mathbf{L}(\Omega_2)$, $h(\hat{1^1}) = h(\hat{2^1}) = s + r$ and $\phi_1(x) = \phi_2(x),\ \forall x\ (h(x) = r),\ x \in \mathbf{L}(\Omega_1)$. It is asserted that $\mathbf{L}(\Omega_1) \cong_\phi \mathbf{L}(\Omega_2)$, where

$$\phi = \begin{cases} \phi_1, & if\ h(x) \leq r \\ \phi_2, & if\ h(x) \geq r. \end{cases}$$

Let us consider two elements $x, y \in \mathbf{L}(\Omega_1),\ h(x), h(y) \geq r$ but $h(x \wedge y) < r$. Let the elements be associated with the generated principal filters $[x)$, $[y)$ and $([x) \vee [y))$, since the join of the two principal filters is the principal filter of the meet of the generating elements $[x) \vee [y) = [x \wedge y)$. On the other hand $[x \wedge y) = \{z : z \geq (x \wedge y)\}$ and $z \leq z'\ if\ z \vee z' = z'\ and\ z \wedge z' = z$. If the lower partial isomorphism is satisfied for all elements $\{z : z \in [x \wedge y),\ h(z) \leq r\}$, as well as the upper partial isomorphism is satisfied for all elements $\{z : z \in [x \wedge y),\ h(z) \geq r\}$ and a bijective mapping ϕ exists, which satisfies the stated condition, then the unique coupling of the elements of filters implies that the isomorphism condition fulfills for x, y and $x \wedge y$. Similar considerations based on the associated ideals $((x], (y]$ and $(x \vee y])$ prove the isomorphism, if (x, y) belong to the lower partial lattice, but their join belongs to the upper partial lattice $\phi(x \vee y) = \phi(x) \vee \phi(y),\ h(x), h(y) \leq r,\ h(x \vee y) \geq r$. □

Lemma 11.

Proof. The lattices $\mathbf{L}(\overline{E})$ and $\mathbf{L}(\overline{F})$ can be considered for products of chain lattices given over the components with elementary operations specified in Definition 12.. Since these chain lattices are distributive, their Cartesian products have the same property. □

Theorem 3.

Proof. Since $\mathbf{L}(\Omega_1) \cong_\phi \mathbf{L}(\Omega_2) \Longleftrightarrow \langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(S_2), \leq \rangle$, the following assertion is proved,

$$\langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(S_2), \leq \rangle \Longleftrightarrow \mathbf{L}(\overline{E}_1) \cong_\phi \mathbf{L}(\overline{E}_2)$$

implying $\mathbf{L}(\overline{E}_1) \cong_\phi \mathbf{L}(\overline{E}_2) \iff \mathbf{L}(\overline{F}_1) \cong_\phi \mathbf{L}(\overline{F}_2)$ because of the construction of involved lattices. Let us consider the following sets

$$E_1' = \{\vec{\alpha}_{\hat{x}/{}^1\hat{0}} \ : \ x \in SC(x/{}^1\hat{1})\}, \quad E_2' = \{\vec{\alpha}_{\phi(\hat{x})/{}^2\hat{0}} \ : \ \phi(x) \in SC(x/{}^2\hat{1})\}$$

Every vector $\vec{e}_1 \in \overline{E}_1$ and $\vec{e}_2 \in \overline{E}_2$ is of the form $\vec{e}_1 = \vec{t}_1 \vee ... \vee \vec{t}_k$, $\vec{e}_2 = \vec{u}_1 \vee ... \vee \vec{u}_s$, where $\vec{t}_1, ..., \vec{t}_k \in E_1'$ and $\vec{u}_1, ..., \vec{u}_s \in E_2'$. Because of the construction of the closure of sets

$$\mathbf{L}(\overline{E}_1') \equiv \mathbf{L}(\overline{E}_1), \quad \mathbf{L}(\overline{E}_2') \equiv \mathbf{L}(\overline{E}_2)$$

and the bijections $\phi_1 : \tau(S_1) \longleftrightarrow E_1'$ and $\phi_2 : \tau(S_2) \longleftrightarrow E_2'$ coupling images of words and generator vectors of sets $\overline{E}_1$ and $\overline{E}_2$

$$\langle \tau(S_1), \leq \rangle \cong_\phi \langle \tau(S_2), \leq \rangle \Longleftrightarrow \mathbf{L}(\overline{E}_1') \cong_\phi \mathbf{L}(\overline{E}_2')$$

proving the statement. The second assertion fulfills by definition of vector lattice operations generating the closure set in an order-preserving way. □

Lemma 12.

Proof. Conditions (14) and (16) are fulfilled trivially. As to the condition (15)

$$\vec{a} \wedge \vec{b} = \hat{0}, \quad \vec{a}, \vec{b} \in \overline{E} \Longrightarrow \big(((a_i = 0 \Longrightarrow b_i \neq 0) \bigwedge (b_i = 0 \Longrightarrow a_i \neq 0).$$

Taking into account the scalar product structure of measure μ, also (15) is ensured. □

Lemma 13.

Proof. For all $\vec{a}(\equiv {}^1\vec{\beta}_{\hat{x}}/\lambda_p) \in \mathbf{L}(\overline{F}_1)$, $\vec{b}(\equiv {}^2\vec{\beta}_{\hat{y}}/\lambda_q) \in \mathbf{L}(\overline{F}_2)$, $\vec{c}(\equiv {}^3\vec{\beta}_{\hat{z}}/\lambda_r) \in \mathbf{L}(\overline{F}_3)$, vectors and associated lattices

$$\mu(\kappa(\vec{a} - \vec{b})) = \mu(\kappa(\vec{b} - \vec{a})) \ \geq \ 0 \ = \ \mu(\kappa(\vec{a} - \vec{a})) \ = \ \mu(\kappa(\vec{b} - \vec{b}))$$

$$\mu(\kappa(\vec{a} - \vec{b})) + \mu(\kappa(\vec{b} - \vec{a})) \ \geq \ \mu(\kappa(\vec{a} - \vec{c})).$$

The last two elations are quite obvious because the measure provides the usual Euclidean distance of shape property vectors. □

7. Appendix 2. (Basics of lattice theory)

The content of this section can be found in many textbooks of algebra or lattice theory [20-24]. It is included here only for convenience, therefore the statements are given without formal proofs. The forthcoming discussion refers to finite lattices. It is more general than our actual topic and the notation will distinctly differ from the previous symbolism.

Lattice theory, as a branch of algebra refers to a set endowed with some operations. The ordered pair $\langle L, \Gamma \rangle$ specifies a lattice, if L is a non-empty set and $\Gamma = \{\gamma_i : i \in I\}$ is an ensemble of operations obeying some conditions. Set L is the 'universe' of the (lattice) algebra and $\gamma_i \in \Gamma,\ (i \in I)$ are the basic operations in it. The operations can be of rank $0, 1, 2$ and they are functions from the product sets L^0, L^1, L^2 into L. Set L^0 is identified with $\{\emptyset\}$, therefore the rank 0 operation selects only a special element of L, but in the actual report only operations of rank two appear explicitly. By a convention we shall use boldface letters ($\mathbf{L}$) to denote algebras and simple uppercase letters (L) for their universes. Lattices can be seen equivalently as special (partially) ordered sets or binary (rank 2) algebras. To be a (partially) ordered set, the (order) relation must be,

- i.) reflexive, i.e. $\forall b(b\rho b)$
- ii.) anti-symmetric, i.e. $b\rho b',\ b'\rho b \implies b = b'$
- iii.) transitive, i.e. $b\rho b',\ b'\rho b'' \implies b\rho b''$.

If considering lattices as binary algebras with operations denoted by $\vee$ and $\wedge$, the following identities must be satisfied, $x, y, z \in L$:

- $\lambda 1.)$ $x \wedge x = x;\ x \vee x = x$ (idempotency).
- $\lambda 2.)$ $x \wedge y = y \wedge x;\ y \vee x = x \vee y$ (commutativity).
- $\lambda 3.)$ $x \wedge (y \wedge z) = (x \wedge y) \wedge z;\ x \vee (y \vee z) = (x \vee y) \vee z$ (associativity).
- $\lambda 4.)$ $x \wedge (x \vee y) = x \vee (x \wedge y) = x$ (absorption law).

If L is a (partially) ordered set with order relation $\leq$, properties i.)-iii.) are obeyed, if every non-void subset $X \subseteq L$ has a least upper bound (l.u.b.), [$a \in L$ is an upper bound of X, if $x \leq a$ for all $x \in X$ and $a \in L$ is a least upper bound of X, if b is also an upper bound of X and $a \leq b$] and a greatest lower bound (g.l.b.), [$a \in L$ is a lower bound of X, if $a \leq x$ for all $x \in X$ and $a \in L$ is a greatest lower bound of X, if b is also a lower bound of X and $b \leq a$]. The lattice operations and order relations are consistent in the following way,

$$a \leq b,\ if\ a = a \wedge b\ and\ b = a \vee b.$$

The lattice operations assign to each couple of elements their l.u.b. (supremum, $\vee$) and g.l.b. (infimum, $\wedge$). The lattice is complete if every subset has a g.l.b and a l.u.b.. The least element is the $\hat{0}$, the greatest is the $\hat{1}$. An element b covers a, if $a < b$ and $a \leq c \leq b \Longrightarrow a = c$ or $c = b$. The upper covers of the $\hat{0}$ are the 'atoms' of the lattice, while elements exactly below $\hat{1}$ are the 'dual atoms'. Elements a, b are comparable, whenever $a \leq b$ or b$\leq a$ and incomparable $(a||b)$, otherwise. Those elements, which are comparable pairwise form a chain, while those which are incomparable form an anti-chain. Using the covering relation, finite lattices can be displayed by drawing a Hasse diagram. Here the elements of L are represented by points on a plane, where each point is connected to the (point-)representatives of upper and lower cover elements. Some subsets of L are closed with respect to one or both of the lattice operations. If U is non-empty, $U \subseteq L$ and for all $b \in L$, if $b \leq a \Longrightarrow b \in U$, furthermore $a, b \in U \implies (a \vee b) \in U$, then U is an 'ideal' of the lattice denoted by $(U]$. By dualization we arrive at the concept of the 'filter' of the lattice. If U is non empty, $U \subseteq L$ and for all $b \in L$, if $b \geq a \Longrightarrow b \in U$, furthermore $a, b \in U \implies (a \wedge b) \in U$, then U is a filter (dual ideal) of the lattice denoted by $[U)$. A sublattice is closed with respect to both of the operations, i.e. if U is non-empty, $U \subseteq L$ and $(a, b) \in U \Longrightarrow (a \vee b) \in U$, $(a \wedge b) \in U$, then U is a sublattice. The sets of ideals (filters, sublattices) can be ordered by the set-theoretic inclusion giving the ideal lattice $\mathbf{I}(L)$.

The morphism concept covers mappings from a lattice into a lattice (possible the same one). Let $\phi : L1 \rightarrow L2$ be a function from $L1$ to $L2$. The mapping is isotone, if $x \leq y \Longleftrightarrow \phi(x) \leq \phi(y)$. It is a meet morphism, if $\phi(x \wedge y) = \phi(x) \wedge \phi(y)$, a join morphism, if $\phi(x \vee y) = \phi(x) \vee \phi(y)$ and a lattice (homo)morphism, if both properties are fulfilled. The mapping is an isomorphism, if it is a bijection, an epimorhism, if it is onto, an endomorphism, if it is a homomorphism and $L1 = L2$ and an automorphism, if it is isomorphism with $L1 = L2$.

A congruence relation is a special kind of equivalence relation. An equivalence relation exhibits similar properties as i.)-iii.), except ii.), which is replaced by

- ii'.) symmetric, i.e. $b \rho b' \implies b' \rho b$.

An equivalence relation ρ is a congruence relation, if the so called 'substitution property' is fulfilled, $x_0 \equiv y_0 \ (\mod \rho)$, $x_1 \equiv y_1 \ (\mod \rho) \implies x_0 \wedge x_1 \equiv y_0 \wedge y_1 \ (\mod \rho)$, $x_0 \vee x_1 \equiv y_0 \vee y_1 \ (\mod \rho)$, where $x \rho y$ was substituted with the notation $x \equiv y (\mod \rho)$. All congruences of a lattice $\mathbf{L}$ can be ordered by inclusion to form the congruence lattice $\mathbf{Con}(\mathbf{L}(L)) =< Con(\mathbf{L}(L)), \wedge, \vee >$, where the meet is the set-theoretic intersection and the join is the transitive closure [22].

The lattices can be classified by those identities, which their elements satisfy. The 'modular' lattices are characterized by the modular identity,

$$x \wedge (y \vee z) = (x \wedge y) \vee z, \ \forall (x, y, z) \in L, \ z \leq x.$$

In the class of 'distributive' lattices, the modular law appears in unrestricted form,

$$x \wedge (y \vee z) = (x \wedge y) \vee (x \wedge z), \ \forall (x, y, z) \in L.$$

The Boolean lattices are distributive, contain the universal bounds $\hat{0}$ and $\hat{1}$ and every element has a unique complement x', with the following properties $x \wedge x' = \hat{0}, \ x \vee x' = \hat{1}$.

We do not intend to pick out further elements of lattice theory, this short overview hopefully covers the topic we are interested in.

References

[1] S.G. Oliver, *From DNA sequence to biological function*, Nature **379**, (1996), pp.597-600.

[2] M. van Heel, *A New Family of Powerful Multivariate Statistical Sequence Analysis Techniques*, J. Mol. Biol. **220**, (1991), pp.877-887.

[3] K. Ohshima, N. Okada, *Generality of tRNA Origin of Short Interspersed Repetitive Elements (SINEs) Characterization of Three Different tRNA-derived Retroposons in the Octopus*, J. Mol. Biol. **243**, (1994), pp.25-37.

[4] S. Steinberg, D. Gautheret, R. Cedergren, *Fitting the Structurally Diverse Animal Mitochondrial tRNAs^{Ser} to Common Three-Dimensional Constraints*, J. Mol. Biol. **236**, (1994), pp.982-989.

[5] E. Hamori, G. Varga, *DNA Sequence (H) Curves of the Human Immunodeficiency Virus 1 and Some Related Viral Genomes*, DNA, **7**, (1988), pp.371-378.

[6] E. Hamori, J. Ruskin, *H Curves, A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences*, The J. of Biol. Chem., **258**, (1983), pp.1318-1327.

[7] D.E. Finkel, *Fractal Displays of Genomic DNA. I. Eco RI Fractal Lattice of Buffalo Rat*, Int. Journ. of Quant. Chem., **36**, (1989), pp.575-586.

[8] G.L. Findley, S.P. McGlynn, *A Generalized Genetic Code*, Int. Journ. of Quant. Chem., Quant. Biol. Symp. **6**, (1979), pp.313-327.

[9] G.L. Findley, S.P. McGlynn, *The Generalized Genetic Code, A Modification of Code Iniversality*, Int. Journ. of Quant. Chem., Quant. Biol. Symp. **7**, (1980), pp.277-281.

[10] A.M. Findley, G.L. Findley, *The Generalized Genetic Code, A Note on Order-Isomorphism/ Order-Equivalence Relations*, Int. Journ. of Quant. Chem., Quant. Biol. Symp. **9**, (1982), pp.59-63.

[11] G.L. Findley, S.P. McGlynn, *Geometry and Evolution*, Int. Journ. of Quant. Chem., Quant. Biol. Symp. **8**, (1981), pp.455-461.

[12] F. Wattenberg, *Nonstandard analysis and the theory of shape*, Fund. Math. **98**, (1978), pp.41-60.

[13] Y. Kodama, S. Spiez, T. Watanabe, *On the shape of hyperspaces*, Fund. Math. **100**, (1978), pp.71-88.

[14] M.A. Johnson, G.M. Maggiora (eds.), *Concepts and Applications of Molecular Similarity*, Wiley&Sons Inc., New York, (1990).

[15] P.G.Mezey, *Shape in Chemistry*, VCH Publishers, New York (1993).

[16] S.Mardesic, J.Segal, *Shape Theory*, North- Holland Publ. Co.,(1982).

[17] P.D. Walker, P.G. Mezey, *Toward Similarity Measures for Macromolecular Bodies*, J. Comp. Chem. **16**,(1995),pp.1238-49.

[18] G.A. Arteca, P.G. Mezey, *Algebraic Approaches to the Shape Analysis of Biological Macromolecules*, in Computational Chemistry: Structure, Interactions and Reactivity, Part A, S. Fraga (ed.), Elsevier, Amsterdam (1992).

[19] G. Náray-Szabó, P.R. Surján, *Computational Methods for Biological Systems* in Theoretical Chemistry of Biological Systems, in Studies in Physical and Theoretical Chemistry, **41**, G. Náray-Szabó (ed.), Elsevier, Amsterdam (1986), pp.1-100.

[20] G.Birkhoff, *Lattice Theory*, American Mathematical Society, New York (1948).

[21] P. Crawley, R.P. Dilworth, *Algebraic Theory of Lattices*, Prentice-Hall Inc., Englewood Cliffs (1973).

[22] R.N. McKenzie, G.P. McNulty, W.F. Taylor, *Algebras, Lattices, Varieties*, Wadsworth&Brooks/Cole, Monterey (1990).

[23] H.H. Schaefer, *Banach Lattices and Positive Operators*, Springer, New York, (1974).

[24] G. Grätzer, *General Lattice Theory*, Academic Press, New York, San Francisco (1976).
[25] I. Bálint, A. Kocsor, I. Cserpán, K. Fodor, to be published.

Department of Theoretical Physics, József Attila University,
Szeged, Hungary, H-6720, Tisza L. krt. 80-82.
Department of Pharmaceutical Analysis, Szent-Györgyi Albert Medical University,
Szeged, Hungary, H-6720, Somogyi Béla u. 4.
Department of Natural Scienses, College of the Technical University of Miskolc,
Dunaújváros, H-2400, Táncsics M. u. 1.
E-mail address: ibalint@sol.cc.u-szeged.hu

Research Group on Artificial Intelligence, József Attila University,, Szeged, Hungary, H-6720,
Aradi vértanúk tere. 1.
E-mail address: kocsor@sol.cc.u-szeged.hu

Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary, H-6701,
PoB:521.
E-mail address: cserpani@everx.szbk.u-szeged.hu

Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary, H-6701,
PoB:521.
E-mail address: fodork@everx.szbk.u-szeged.hu