International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:7, No:12, 2013

A Geometrical Perspective on the Insulin Evolution

Yuhei Kunihiro, Sorin V. Sabau, Kazuhiro Shibuya

Abstract—We study the molecular evolution of insulin from metric geometry point of view. In mathematics, and in particular in geometry, distances and metrics between objects are of fundamental importance.

Using a weaker notion than the classical distance, namely the weighted quasi-metrics, one can study the geometry of biological sequences (DNA, mRNA, or proteins) space.

We analyze from geometrical point of view a family of 60 insulin homologous sequences ranging on a large variety of living organisms from human to the nematode *C. elegans*. We show that the distances between sequences provide important information about the evolution and function of insulin.

Keywords—Metric geometry, evolution, insulin.

I. Introduction

FOR a given set X, the classical notion of **distance**, or **metric**, is a mapping $d : X \times X \to [0, \infty)$ which is non-negative, symmetric and satisfies the triangle inequality (see [3] for details). The **quasi-metrics** are generalizations of classical metrics in the sense that they do not have anymore the symmetry property, i.e. a mapping $d : X \times X \to [0, \infty)$ such that for $\forall x, y \in X$, we have $d(x, y) = d(y, x) = 0 \Leftrightarrow x = y$ and triangle inequality $d(x, y) + d(y, z) \geq d(x, z)$ for $\forall x, y, z \in X$.

Quasi-metric spaces (X, d) have several geometric structures. Indeed, we can introduce the **reverse** quasi-metric $\bar{d}(x, y) := d(y, x)$ and the **symmetrization** of d given by $\rho(x, y) = \frac{1}{2}[d(x, y) + d(y, x)]$, for $\forall x, y \in X$. In the metric geometry ρ, d and $\bar{d}$ are called the **average**, **forward** and **backward** distance on X, respectively.

A special case is the so-called **weighted quasi-metric** (d, w), where d is a quasi-metric and $w : X \to [0, \infty)$ is a mapping, satisfying $d(x, y) + w(x) = d(y, x) + w(y)$, for $\forall x, y \in X$, called the **weight** of d.

The weighted quasi-metrics were initially introduced in the context of theoretical computer science ([12]) and their topological properties were extensively studied ([11]). Recently, it has been shown that weighted quasi-metrics are essential for sequence comparison in molecular biology and bioinformatics ([15]). Indeed, it is shown in [15] that the evolutionary distances induced by the PAM and BLOSUM similarity scores are actually weighted quasi-metrics defined on the free monoids with finite generators of the nucleotides A, C, T, G or the 20 proteinogenic amino acids $A, V, L, I, P, M, F, W, G, S, T, C, Y, N, Q, D, E, H, K, R$.

These type of quasi-metrics have a strong mathematical motivation because if X is a smooth manifold, then these correspond to a certain class of Finsler metrics ([14]).

The geometrical properties of such evolutionary metrics have also been studied. Indeed, it is known that the PAM-induced metric geometry is inconsistent because the induced distance violates the triangle inequality, but on the other hand, for the BLOSUM score matrices obtained from sequences with more than 40% identity, the induced metric is a well defined weighted quasi-metric on the set of 20 proteinogenic amino acids and therefore the theory of sequence comparison can be modeled as the geometry of a weighted quasi-metric space ([2], [10]). Moreover, we have studied the correlation between the metrical properties of the finite generators set of 20 proteinogenic amino acids and the physical-chemical properties of these amino acids. The hydrophobicity domain was well conserved regardless the metric used. The other properties of amino acids as being aliphatic, aromatic, basic or small were also well identified by the weighted quasi-metrics, while the acidic amino acids D, E and amides N, Q recognition depends on the metric considered ([10]). This study also has shown that the metric geometry of the 20 proteinogenic amino acids space is in accordance with Kimura's neutral theory of evolution ([7]).

In the present paper we extend our investigations from the set of 20 amino acids to a set of insulin homologous sequences.

Insulin and related peptides are essential for growth, development and metabolism. Initially discovered in mammals, the presence of insulin-like peptides in the four major groups of animal kingdom: the chordates and vertebrates; the equinodermos and tentaculatus; the coelenterates; and the molluscs, worms and arthropods is nowadays generally accepted ([5]). The discovery that brain produced insulin-like peptides in insects and molluscs are involved in the control of growth shows that the role of insulin is not confined to glucose metabolisms, but also related to growth.

The studies of insulin-like protein and associated pathways in the nematode *C. elegans* have shown that the genetic mechanism of regulating glucose metabolism in mammals is evolutionary very ancient ([6]).

Generally, in evolution, the insulin is considered to be initially a neurohormone produced by brain and responsible for both growth and metabolism regulation. However, because the glucose metabolism in vertebrates is different from insects, the insulin production has been changed from brain to pancreas in order to achieve an efficient regulation of glucose levels in blood. All these show that the action of insulin to

Yuhei Kunihiro is with Tokai University, School of Science and Engineering, 5-1-1-1, Minami-Sawa, Minami-ku, Sapporo 005-8601, Japan (e-mail: 2bsim001@mail.tokai-u.jp).

Sorin V. Sabau is with Tokai University, School of Science, Department of Mathematics, 5-1-1-1, Minami-Sawa, Minami-ku, Sapporo 005-8601, Japan (phone: +81-11-571-5111, fax: +81-11-571-7879, e-mail: sorin@tspirit.tokai-u.jp).

Kazuhiro Shibuya is with Hiroshima University, Department of Mathematics, 1-7-1, Kagamiyama, Higashi-Hiroshima, 739-8511, Japan (e-mail: shibuya@hiroshima-u.ac.jp).

regulate development, growth or reproduction is more ancient than its action to regulate blood glucose levels ([8]).

This rather unexpected structural and functional evolution of insulin makes it a very interesting research model not only for the evolution of life on Earth, but also for the paradoxes it begins with ([8]).

We show that there is a very good correlation between the metric geometry of the insulin homologous sequences and the evolution of this hormone-peptide. Moreover, we study some biological properties of insulin from the geometric point of view.

A comparison of the primary amino acid sequence of insulin from a wide range of mammalian, non-mammalian vertebrates to the nematode *C. elegans* proves the existence of a strong evolutionary tendency to conserve those amino acids that are essential for the insulin function, i.e. binding to its receptor. We have evaluated this from a geometrical point of view.

II. Sequences, similarities, distances

We recall here the fundaments of sequence comparison geometry and relation with similarity (see [2], [10], [15] for details).

Let $\sum$ be a finite dimensional non-empty set, called the **alphabet**. The elements $u_1 \dots u_n \in \sum$ are called **letters** of the alphabet $\sum$. We consider the free **monoid** $\sum^*$ over $\sum$, namely the set of finite sequences of zero or more elements from $\sum$, whose elements $u_1^* \dots u_k^*$ are called **words** or **strings**, while the letters in the alphabet are called **generators**. The operation of the monoid $\sum^*$ is the binary operation of concatenation. The empty string $\{e\}$ is the identity element. It is customary to denote by $\sum^+$ the subset of $\sum^*$ containing all elements except the identity. $\sum^+$ is a free semigroup over the generators set $\sum$.

There is a strong biological motivation to consider these mathematical notions. Indeed, the macromolecules that contain the fundamental information of life can be expressed as words over a finite alphabet. For example, **DNA molecules** are words in the free semigroup generated by the four letters **nucleotide alphabet** (A, C, T, G), **RNA molecules** are words in the free semigroup generated by the four letters **nucleotide alphabet** (A, C, U, G), and **proteins molecules** are words in the free semigroup generated by the **proteinogenic amino acid alphabet** $(A, V, L, I, P, M, F, W, G, S, T, C, Y, N, Q, D, E, H, K, R)$, where each amino acid is denoted by a letter.

A fundamental example is the insulin and pre-proinsulin that have facilitated our understanding of molecular evolution. It is known that the human pre-proinsulin molecule consists of signal peptide (required for intracellular transport; amino acid residues 1-24), the B chain, the C peptide, and the A chain. Dibasic residues (RR, KR) flank the C peptide and are the sites at which proteases cleave the protein. The A chain and B chain are then covalently linked through disulfide bridges, forming mature insulin (see Fig. 1).

Fig. 1 The human preproinsulin molecule structure

Formally, we can write proinsulin:=$u_1^*\ u_2^*\ u_3^*\ u_4^*$, where $u_1^*.u_2^*, u_3^*, u_4^*$ are words over the 20 proteinogenic amino acids alphabet. Namely, we have:

- $u_1^* = MALWMRLLPLLALLALWGPDPAAA$
- $u_2^* = FVNQHLCGSHLVEAL$ $YLVCGERGFFYTPKT$
- $u_3^* = (RR)EAEDLQVGQVE$ $LGGGPGAGSLQPLALEGSLQ(KR)$
- $u_4^* = GIVEQCCTSICSLYQLENYCN$

Sequence comparison is one of the most important research areas in bioinformatics and molecular theory of evolution. The classical and most basic tool used is the NCBI BLAST which, for a given DNA or protein sequence, retrieves all similar sequences from a sequence database. The BLAST algorithm is computing the similarities between sequences using **alignments**. There are two types of sequence comparison: **global**, between whole sequences, and **local**, between fragments of sequences.

The similarity scores on nucleotides or amino acids, as well as the gaps penalties introduced in sequences have both biological and statistical interpretation. More precisely, let $s : \sum \times \sum \to \mathbb{R}$ be a similarity score matrix, namely a mapping having the properties $s(x, x) \geq 0$, $s(x, y) = s(y, x)$ and $s(x, x) \geq s(x, y), \forall x, y \in \sum$. It is known that score matrices can be converted into distances between amino acids by $d(x, y) := s(x, x) - s(x, y)$. Since $s(x, x) \neq s(y, y)$ in general it follows $d(x, y) \neq d(y, x)$ and therefore $(\sum, d)$ is a quasi-metric space ([2], [10], [15]).

The main relation of bioinformatics with geometry is given by the fundamental observation that maximizing sequence similarities is equivalent to minimizing distances.

In order to compare sequences, except distances between amino acids one needs **string edit distances**, i.e. the smallest number of permitted edit operations required to transform one string into another. Here permitted edit operations are: substitutions, insertions, and deletions.

Mappings that transforms a sequence into another are called **alignments** and these can be computed manually for short sequences or by means of dynamic programming algorithms for long sequences. There are **global alignments** where sequences are aligned in their full lengths (for eg. Needleman-Wunsch algorithm) and **local alignments** where sub-sequences are globally aligned (for eg. Smith-Waterman algorithm, see for e.g. [13]).

Based on the score matrices PAM or BLOSUM one can define weights and gap-penalties for alignments of sequences and from here string edit distances between aligned sequences.

In some cases, local alignments and induced local similarities, or equivalently, local distances, are preferred for comparison of biological sequences.

Indeed, the sequences carrying biologically significant information are usually discrete short fragments of sequences.

Strong similarities of such segments do not extend to a similarity of full sequence. For example, the function of a protein induced by its 3D structure is determined by discrete structural domains. Hence, even functionally closely related protein sequences may show little similarity outside the conserved domains and their global similarity may not be statistically significant ([15]). However, in the case of insulin homologous sequences comparison, since all sequences have some function we prefer the global alignment.

III. Geometry of Insulin Molecular Evolution

As mentioned already, the study of insulin has contributed substantially to the understanding of molecular evolution of life on Earth. The mature, biologically active protein consists of two subunits, the A chain and B chain, that are covalently attached through intermolecular disulphide bridges ([13]).

Earlier evolutionary studies have started by comparing few species (cow, sheep, pig, horse, whale) and it became immediately clear that at least for vertebrate organisms of the class Mammalia the A chain and B chain residues are highly conserved and the amino acid substitutions were restricted to three residues within a sulphide loop region of the A chain. These observations suggest that amino acid substitutions occur non-randomly, some changes could alter dramatically the biological function of the protein, while others can be neutral. The rate of these changes are in accord with Kimura's neutral theory of evolution ([13]), see Fig. 2.

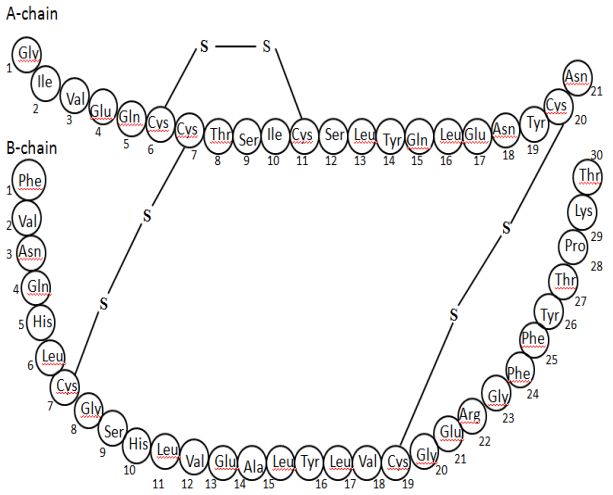

Fig. 2 The structure of human insulin

We have performed a quick search for insulin in the NCBI database and extract from the hits 91 insulin homologous sequences belonging to 73 different organisms. After performing the alignment of these sequences by ClustalX (http://www.clustal.org/), it can be seen that some evolutionary closed organisms have identical insulin sequences. We have built in this way a set of 60 non-redundant insulin homologous sequences (see Table VI).

We remark that human insulin (Acc. No. NP_000198) has identical primary amino acid sequence with the primates chimpanzee (*Pan troglodytes*), Savanna Monkey (*Chlorocebus aethiops*), and Crab-eating macaque (*Macaca fascicularis*) (Acc. No. P30410, CAA43405, and AAA36849, respectively).

Based on these alignments we have computed the evolutionary distances between these 60 sequences by using ClustalX distance, the local alignment induced forward, backward and average distances (data not shown). The local forward and backward distances were computed using the software in EMBOSS web site (http://emboss.sourceforge.net/). Moreover have constructed rooted and unrooted phylogenetic trees correlated with hierarchical clustering methods in order to infer the evolutionary relations between these sequences using the joint neighbour method and NJplot (http://pbil.univ-lyon1.fr/software/njplot.html) for graphical representation (data not shown). Recall that the forward and backward distances are not symmetric while ClustalX and averaged distances are symmetric.

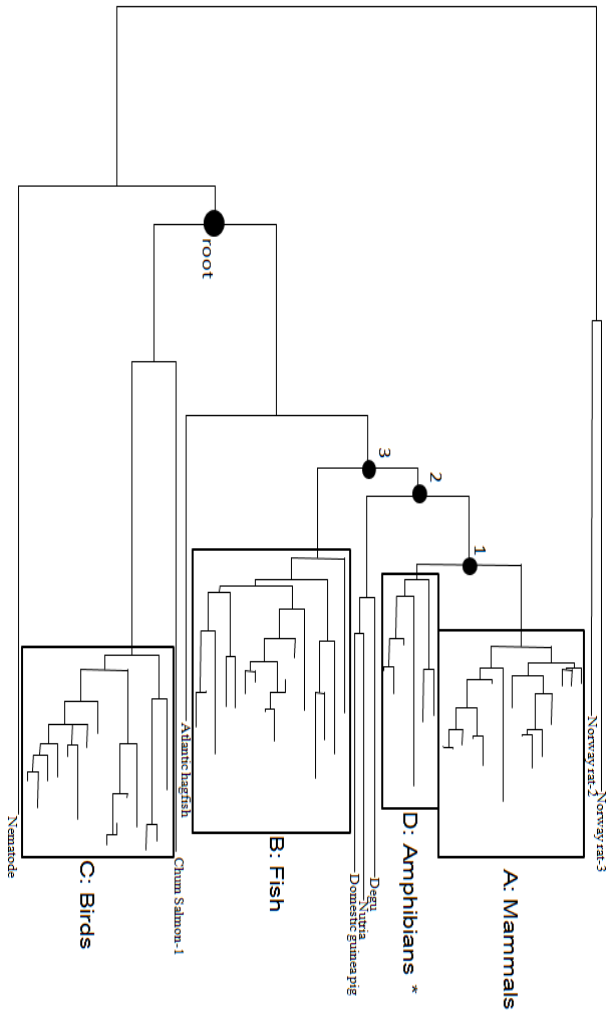

Fig. 3 A rooted phylogenetic tree obtained from ClustalX distances

TABLE I
MAMMALS

Number	Organism name
1	Human, Chimpanzee, Savanna Monkey, Crab-eating macaque
2	Northern Night Monkey, Common Squirrel Monkey
3	Mouse-1(Formosan field), Ryukyu mouse-1, House mouse-1, Coxing's white-bellied rat-1
4	Mouse-2(Formosan field), Lesser rice-field rat-1, House mouse-2, Coxing's white-bellied rat-2, Norway rat-1
5	Mongolian gerbil rat, Fat sand rat
6	Golden Hamster, Taiwan vole
7	Ryukyu mouse-2
8	House mouse-3
9	Lesser rice-field rat-2
14	Boar, Dog, Domestic pig
15	Cattle
16	Cat
17	European rabbit
18	Sheep
20	Elephant

TABLE II
FISH

Number	Organism name
39	Catla, Common Carp
40	White Sucker
41	Zebrafish
42	Goldeye
43	Silver arowana
44	Barfin flounder (flot)
45	Knifefish
46	Monkfish
49	Chum Salmon-2
50	Salmon_and_rainbow trout (Salmon-7, Rainbow Trout)
51	Salmon-3
52	Salmon-4
53	Salmon-5
54	Salmon-6
55	Butterflyfish
56	Cod

TABLE III
BIRDS

Number	Organism name
19	Emperor Penguin, Northern fulmar, Common Moorhen, Black-crowned night heron, Snow Petrel, Black-legged kittiwake, Red-footed booby, Common Murre
22	Chicken-1
25	European Green-wing Teal, Mallard
26	Muscovy Duck, Shelduck
27	White Stork
28	Rock pigeon-1
30	Emu, African Ostrich, Greater Rhea
31	Eurasian Kestrel
32	Turkey-1
33	Budgerigar
34	Domestic Sparrow
35	Tropicbird
36	Common Magpie, Blackbird
37	Tawny Owl
38	Black-browed Albatross

TABLE IV
AMPHIBIANS*

Number	Organism name
23	Chicken-2, Ostrich, Turkey-2
24	Rufous hummingbird
29	Rock pigeon-2
57	Leopard Frog
58	African Clawed Toad-1
59	African Clawed Toad-2

TABLE V
OTHERS

Number	Organism name
10	Norway rat-2
11	Norway rat-3
12	Domestic guinea pig
13	Nutria
21	Degu
47	Atlantic hagfish
48	Chum Salmon-1
60	Nematode

The rooted tree obtained from the averaged distance matrix identifies very well the outgroups of the nematode *C. elegans* insulin-like sequence (Acc. No. AAC33275) and the insulin receptor sequences Norway rat-2 (Acc. No. AAB38968) and Norway rat-3 (Acc. No. AAB38967) (Fig. 3). The root can be now easily inferred and further one can easily identify 4 clades A, B, C, D that contain evolutionary closely related sequences (see Table I $\sim$ IV for the members of each clade).

The group D (see Table IV) contains some birds and amphibians and a quick look at the insulin homologous sequence alignment Table VI shows the striking similarity

between these.

The internal node 1 (Fig. 3) represents an inferred ancestor sequence of mammals and amphibians* groups. The internal node 2 represents an inferred ancestor sequence of mammals, amphibians* a special group of insulin homologous sequences belonging to degu, nutria, and guinea pig.

We remark that degu, nutria and guinea pig insulin sequences are closely represented. The organism themselves are evolutionary close and have in common the property that, unlike other types of insulin, their insulin do not bind two zinc ions (see Table VI).

The internal node 3 represents an inferred ancestor sequence of mammals, amphibians* and fish insulin sequences clades A, B, D (see Table I, II, IV). Finally the root represents an inferred ancestor sequence of all insulin sequences present in the analysis.

From geometric point of view, the amphibians* insulin homologous sequences are closer to the mammals sequences than the fish and birds analogous sequences, respectively. This shows that a metric geometry approach to evolution can reveal facts that one hidden when one based mainly on palaeontological data. Further investigations are necessary in order to clarify the biological mechanism underlying this fact.

A similar study can be done for the phylogenetic trees obtained from the forward and backward distances.

We point out that regardless the metric used the topology of the phylogenetic trees do not changes essentially. There are however very small variations.

We will consider now the clade A of organisms (see Table I). We remark that, regardless the metric used, the human insulin (Acc. No. NP_000198) is most closely related to the European rabbit insulin sequence (Acc. No. NP_001075804) found at only 3.5, 4 and 3 units for averaged, forward and backward metric, respectively.

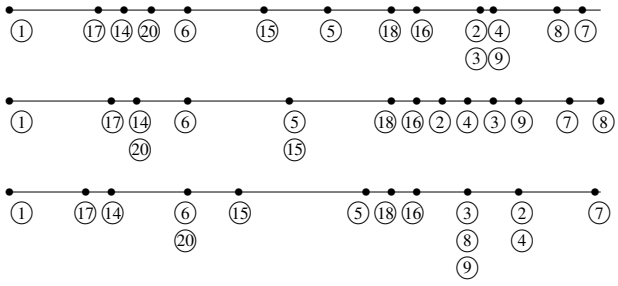

Fig. 4 The order of organisms in clade A using the average, forward and backward distance (from top to bottom). The numbers used here are the numbers in first columns in Table I

Indeed, the alignment in Table VI shows that the only difference between primary amino acid sequence of human and European rabbit is the substitution SerB30 $\rightarrow$ ThrB30. Both Serine (Ser) and Threonine (Thr) are small, hydrophilic, polar uncharged amino acids (with evolutionary distance of 3.5, 4 and 3 units with respect to the averaged, forward and backward evolutionary BLOSUM62-distance, respectively) so this substitution do not affects essentially the function of the insulin molecule. Therefore, the small distance is indeed motivated by the biological properties of the protein.

The European rabbit is followed by the boar, dog, pig insulin sequence found at 4.5, 5 and 4 units from human insulin with respect to the averaged, forward and backward BLOSUM62 induced evolutionary distance, respectively (see Fig. 4). Indeed, alignment in Table VI shows the substitution AlaB30 $\rightarrow$ ThrB30. Unlikely the European rabbit case, threonine and alanine are hydrophilic and hydrophobic amino acids. respectively, with different physico-chemical properties, fact that motivates the larger evolutionary distance.

It is known that the insulin molecule in vertebrate mammals associates into stable dimers and, in the presence of Zn^{2+} ions, forms stable hexamers. Past research suggests that the receptor-binding region of insulin is related to the dimer-forming surface and it is made of the amino acids found at positions; ValB12, TyrB16, GlyB23-TyrB26, GlyA1-GlnA5, TyrA19, and AsnA21. Residues GlyB20 and ProB28 are also considered to be important on the dimer formation and residues LeuB6, HisB10, AlaB14, LeuB17, ValB18, LeuA13 and TyrA14 are involved in the hexamer formation (see [4] and references in [1]). An alternative model is that the receptor binding domain consists of only five residues: IleA2, ValA3, TyrA19, GlyB23, PheB24 forming a patch on the surface of the molecule. Likely residues LeuB6, GlyB8, LeuB11, GluB13 and PheB25, even though not related to the insulin binding site one of importance in the conformation of the insulin (see [4], [9]).

A quick look at the alignments (Table VI) shows that the primates group Northern night monkey, Common squirrel monkey, found at 18.5, 17 and 20 evolutionary units from human insulin with respect to average, forward and backward distance, respectively, differs from human sequence by 5 substitutions amongst which only ValA2 $\rightarrow$ IleA2 and AspA4 $\rightarrow$ GluA4 are related to the receptor binding domain. Valine (Val) and Isoleucine (Ile) are both hydrophobic aliphatic amino acids while Aspartic acid (Asp) and Glutamic acid (Glu) are acidic negatively changed, hydrophilic amino acids.

Hence we can conclude that all insulin sequences in clade A bind to the human insulin receptor. From geometrical point of view, the receptor binding property, on the structure of insulin molecule, potentially extends to the average, forward and backward balls of radius 22.5, 25 and 23 evolutionary units centered at human insulin, respectively.

The next group of organisms close to the clade A is clade D, containing three aviatic and three amphibians insulin homologous proteins. It can be seen that from metric point of view, regardless the metric used, the order from human insulin is: Rufous hummingbird, Chicken-2, Ostrich, Turkey-2, Rock pigeon-2, African Clawed Toad-1, African Clawed Toad-2, Leopard frog. The metric separates efficiently the aviatic insulin from amphibians positioning the first in the closed ball of radius 40.5, 39 and 42 with respect to average, forward and backward metric, respectively, centered at human insulin. The clade D extends to an average, forward and backward distance of 53, 49 and 57 units from human insulin, respectively.

The alignment (Table VI) shows the substitution MetB17

$\rightarrow$ LeuB17 for leopard frog and PheA14 $\rightarrow$ TyrA14 for African Clawed Toad sequences. These substitutions suggests a structural change in the protein structure because methionine (Met) and leucine (Leu) belong to the sulfur containing and aliphatic amino acids group, respectively. Likely, phenylalanine (Phe) and tyrosine (Tyr) are both aromatic and hydrophobic amino acids, but Phe is more hydrophobic than Tyr which is the least hydrophobic aromatic amino acid.

We can conclude that the human insulin receptor binding site structure extends to the aviatic insulin sequences in clade D, but it should be a decreasing in the human insulin-receptor affinity for amphibians insulin sequences. Geometrically, we find the insulin structure well preserved at an average, forward and backward distance of 40.5, 39 and 42 units from the human insulin. After this threshold the receptor binding affinity is thought to decrease.

A similar analysis can be done for the other organisms like Degu, Nutria or Guinea pig (see Table V) and for the organisms in clades B and C (see Tables II, III). It is believed that those organism's insulin have been evolved independently from the human insulin-like sequences.

Finally we discuss the Atlantic hagfish insulin sequence that is known to have only 5% of the potency of human or pig insulin in stimulating biogenesis in isolated rat fat cells. The alignment (Table VI) shown the following substitutions between human and hagfish insulin sequence which related to the binding domain structure: GlyB6 $\rightarrow$ GlnB6, LysB11 $\rightarrow$ SerB11, AspB12 $\rightarrow$ HisB12, AlaB20 $\rightarrow$ ValB20, ValB23 $\rightarrow$ GluB23, IleA13 $\rightarrow$ LeuA13. Among these substitutions, the residues B6 and B23 contains amino acids with different chemical structure and hydrophobicity levels, while those at B11, B12 contains amino acids with different chemical structure. The rest are neutral substitutions.

Therefore it can be concluded that at an average, forward and backward distance of 103, 100 and 106 units from human insulin sequence, the effectiveness of the insulin drops to 5%.

IV. Conclusion

We have studied a set of 60 non-redundant insulin homologus sequences from the metric geometry point of view. These are elements (nodes) of the free monoid $\sum^*$ whose generators set $\sum$ is made of the 20 proteinogenic amino acids. The set $\sum^*$, that is the space of insulin homologous sequences, has a natural structure of weighted quasi-metric space, that is it has the average, forward and backward metric structures, respectively.

Using any of these metric structures one can study the set $\sum^*$ from evolutionary and functional point of view. Even though there are differences when using different metrics, the topology of $\sum^*$ is basically the same fact proving that there is a high degree of consistency between the geometry and topology of the insulin homologous sequences space. Still one should pay attention to small differences when choosing a metric.

The phylogenetic analysis of the space $\sum^*$ clustered our 60 sequences in 5 clades with good biological significance (see Fig. 3, Table I $\sim$ V). Moreover, we have determined that for organisms in clade A, namely in the domain extending to 22.5, 25 and 23 evolutionary units from the human insulin sequence (Acc. No. NP_000198), with respect to the average, forward and backward distance, respectively, the insulin binding site is well preserved and therefore the function is unaltered. The binding site structure seems to be preserved up to 40.5, 39 and 42 evolutionary units from the human insulin, but no further with respect to the average, forward and backward distance, respectively. This corresponds to the insulin homologous sequence in the birds in clade D. Finally, we have shown that at a distance of 103, 100 and 106 evolutionary units from the human insulin sequence, the protein structure changes dramatically, its effectiveness decreasing 5%.

We conclude that metric geometry is a new, still not enough well explored, method to investigate function of proteins form evolutionary point of view that can facilitate our understanding about the birth and evolution of life on Earth.

TABLE VI
THE 60 NON-REDUNDANT INSULIN HOMOLOGOUS SEQUENCES

Number	Organism name	Accession number	Insulin sequence (B chain)	Insulin sequence (A chain)
1	Human, Chimpanzee, Savanna Monkey, Crab-eating macaque	NP_000198(preproinsulin), P30410(preproinsulin), CAA43405(preproinsulin), AAA36849(preproinsulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKT	GIVEQCCTSICSLYQLENYCN——
2	Northern night monkey, Common Squirrel Monkey	AAA35374(preproinsulin), INMKSQ(insulin)	–FVNQHLCGPHLVEALYLVCGERGFFYAPKT	GVVDQCCTSICSLYQLQNYCN——
3	Mouse-1 (Formosan field), Ryukyu mouse-1, House mouse-1, Coxing's white-bellied rat-1	ABB89748(preproinsulin), ABB89749(preproinsulin), CAA28433(preproinsulin), ABB89750(preproinsulin)	–FVKQHLCGSHLVEALYLVCGERGFFYTPMS	GIVDQCCTSICSLYQLENYCN——
4	Mouse-2 (Formosan field), Lesser rice-field rat-1, House mouse-2, Coxing's white-bellied rat-2, Norway rat-1	ABB89744(preproinsulin), ABB89743(preproinsulin), NP_032412(preproinsulin), ABB89746(preproinsulin), ABB89747(preproinsulin)	–FVKQHLCGPHLVEALYLVCGERGFFYTPKS	GIVDQCCTSICSLYQLENYCN——
5	Mongolian gerbil rat, Fat sand rat	ABB89751(preproinsulin), CAA66897(preproinsulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKF	GIVEQCCTGICSLYQLENYCN——
6	Golden Hamster, Taiwan vole	I48166(insulin precursor), ABB89752(preproinsulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKS	GIVDQCCTSICSLYQLENYCN——
7	Ryukyu mouse-2	ABB89745(preproinsulin)	–FVKQHLCGPHLVEALYLVCGERGFFYSPKS	GIVDQCCTSICSLYQLENYCN——
8	House mouse-3	ACX53313 (insulin precursor, partial)	–FVKQHLCGSHLVEALYLVCGERGFFYTPMS	GIVDQCCTSICSLYQLENYC——
9	Lesser rice-field rat-2	ABB89747(preproinsulin)	–FVKQHLCGSHLVEALYLVCGERGFFYTPVS	GIVDQCCTSICSLYQLENYCN——
10	Norway rat-2	AAB38968 (insulin receptor, partial)	—FSVVIGSIYLFLRKRQPDGPMGPLYAPER	—LTDLMRMCWQF————
11	Norway rat-3	AAB38967 (insulin receptor, partial)	——KTFEDYLHNVVFVPRKTSSGNGAPER	—LTDLMRMCWQF————
12	Domestic guinea pig	P01329(preproinsulin)	–FVSRHLCGSNLVETLYSVCQDDGFFYIPKD	GIVDQCCTGTCTRHQLQSYCN——
13	Nutria	720710A(insulin)	–YVSQRLCGSQLVDTLYSVCRHRG-FYRPND	GIVDQCCTNICSRNQLMSYCND—-
14	Boar, Dog, Domestic pig	NP_001103242 (insulin precursor), P01321(preproinsulin), 550086A(insulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKA	GIVEQCCTSICSLYQLENYCN——
15	Cattle	NP_776351(preproinsulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKA	GIVEQCCASVCSLYQLENYCN——
16	Cat	BAB84110(preproinsulin), NP_001009272 (insulin precursor)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKA	GIVEQCCASVCSLYQLEHYCN——
17	European rabbit	NP_001075804 (insulin precursor)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKS	GIVEQCCTSICSLYQLENYCN——
18	Sheep	P01318(preproinsulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKA	GIVEQCCAGVCSLYQLENYCN——

Number	Organism name	Accession number	Insulin sequence(B chain)	Insulin sequence(A chain)
19	Emperor Penguin, Northern fulmar, Common Moorhen, Black-crowned night heron, Snow Petrel, Black-legged kittiwake, Red-footed booby, Common Murre	AAP45980(preproinsulin; C peptide, Achain), AAP45987(preproinsulin; C peptide, Achain), AAP45988(preproinsulin; C peptide, Achain), AAP45992(preproinsulin; C peptide, Achain), AAP45993(preproinsulin; C peptide, Achain), AAP45998(preproinsulin; C peptide, Achain), AAP46001(preproinsulin; C peptide, Achain), AAP46004(preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFQQEEFEKVKR	GIVEQCCHNTCSLYQL———
20	Elephant	INEL(insulin)	–FVNQHLCGSHLVEALYLVCGERGFFYTPKT	GIVEQCCTGVCSLYQLENYCN—
21	Degu	AAA40590(preproinsulin)	–YSSQHLCGSNLVEALYMTCGRSG-FYRPHD	GIVDQCCNNICTFNQLQNYCNVP—
22	Chicken-1	AAP45989(preproinsulin; C peptide, Achain) +AAQ85159 (preproinsulin, partial)	———SSPLRGEAGVLPFQQEEYEKVKR	GIVEQCCHNTCSLYQLMALW——
23	Chicken-2, Ostrich, Turkey-2	CAA41738(preproinsulin), INOS(insulin), 720927A(insulin)	–AANQHLCGSHLVEALYLVCGERGFFYSPKA	GIVEQCCHNTCSLYQLENYCN—
24	Rufous hummingbird	AAC64211 (preproinsulin, partial)	–AVNQHLCGSHLVEALYLVCGERGFFYSPKA	GIVEQCCHNTCSLYQLENYCN—
25	European Green-wing Teal, Mallard	AAP45977(preproinsulin; C peptide, Achain) AAP45979(preproinsulin; C peptide, Achain)	———NGPLHGEVGELPFQHEEYQKVKR	GIVEQCCENPCSLYQL———
26	Muscovy Duck, Shelduck	AAP45981(preproinsulin; C peptide, Achain), AAP46002(preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFQHEEYQKVKR	GIVEQCCENPCSLYQL———
27	White Stork	AAP45982 (preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFQQEEFEKVKR	GIVEQCCHNPCSLYQL———
28	Rock pigeon-1	AAP45983 (preproinsulin; C peptide, Achain)	———SSPLRGEAGVLPFQQEEYEKVKR	GIVEQCCHNTCSLYQ———
29	Rock pigeon-2	EMC88047(preproinsulin)	–AANQHLCGSHLVEALYLVCGDRGFFYSPKA	GIVEQCCHNTCSLYQLENYCN—
30	Emu, African Ostrich, Greater Rhea	AAP45985(preproinsulin; C peptide, Achain), AAP46000(preproinsulin; C peptide, Achain), AAP45997(preproinsulin; C peptide, Achain)	———SGPLRGEAEELPFQQEEYEKVKR	GIVEQCCHNTCSLYQL———
31	Eurasian Kestrel	AAP45986(preproinsulin; C peptide, Achain)	———SSPLHGEAGELPFQQEEFEKVKR	GIVEQCCHNTCSLYQL———
32	Turkey-1	AAP45990(preproinsulin; C peptide, Achain)	———GSPLRGEAGVLPFQQEEYEKVKR	GIVEQCCHNTCSLYQ———
33	Budgerigar	AAP45991(preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFRPEEFQKVKR	GIVEQCCHNTCSLYQL———
34	Domestic Sparrow	AAP45994(preproinsulin; C peptide, Achain)	———SGPLHGELGELPFQQEEFETVKR	GIVEQCCHNTCSLYQL———
35	Tropicbird	AAP45995(preproinsulin; C peptide, Achain)	———SGPLHGEAGELPFQQEEFEKVRR	GIVEQCCHNTCSLYQL———

Number	Organism name	Accession number	Insulin sequence(B chain)	Insulin sequence(A chain)
36	Common Magpie, Blackbird	AAP45996(preproinsulin; C peptide, Achain), AAP46003(preproinsulin; C peptide, Achain)	———SGPLHGELGELPFQQEEFEKVKR	GIVEQCCHNTCSLYQL———-
37	Tawny Owl	AAP45999(preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFHQEEFEKVKR	GIVEQCCHNTCSLYQ———
38	Black-browed Albatross	AAP45984(preproinsulin; C peptide, Achain)	———SGPLHGEVGELPFQQEEFEKVKR	GIVEQCCHSTCSLYQL———-
39	Catla, Common Carp	AAK51558(preproinsulin), 1012233A(preproinsulin)	–GAPQHLCGSHLVDALYLVCGPTGFFYNPKR	GIVEQCCHKPCSIFELQNYCN——
40	White Sucker	AAK28709 (preproinsulin, patrial)	–VAPQHLCGSHLVDALYLVCGPTGFFYNPKR	GIVEQCCHRPCNIFDLEKYCN——
41	Zebrafish	NP_571131(preproinsulin)	—TPQHLCGSHLVDALYLVCGPTGFFYNPK-	GIVEQCCHKPCSIFELQNYCN——
42	Goldeye	AAK54684 (preproinsulin, patrial)	–SSSQHLCGSHLVDALYMVCGEKGFFYQPKT	GIVEQCCHRPCNIFDLNQYCN——
43	Silver arowana	AAK28713 (preproinsulin, patrial)	–SSSQRLCGSHLVDALYMVCGDRGFFYSPKS	GIVEQCCHRPCNIFDLQNYCN——
44	Barfin flounder (flot)	Q9W7R2(preproinsulin)	–LPPQHLCGAHLVDALYLVCGERGFFYTPKR	GIVEQCCHKPCNIFDLQNYCN——
45	Knifefish	AAK28710 (preproinsulin, partial)	–ASNQHLCGSHLVEALYLVCGERGFFYNPKM	GIVEQCCHRPCNIFDQNQYCN——
46	Monkfish	P69045(preproinsulin)	–APAQHLCGSHLVDALYLVCGDRGFFYNPKR	GIVEQCCHRPCNIFDLQNYCN——
47	Atlantic hagfish	P01342(preproinsulin)	–RTTGHLCGKDLVNALYIACGVRGFFYDPTK	GIVEQCCHKRCSIYDLENYCN——
48	Chum Salmon-1	AAA49416 (preproinsulin, partial)	DPLIGFLFPKSSQENEVAEYPFKDQMDMIVKR	GIVEQCCHKPCNIFDLQNYCN——
49	Chum Salmon-2	AAA49417 (insulin B chain, partial) +1412252A(insulin A)	——QLCGSHLVDALYLVCGEKGFFYTPK-	GIVEQCCHKPCNIFDLQNYCN——
50	Salmon_and_ rainbow trout (Salmon-7, Rainbow Trout)	1616203A(insulin), ABN69072(preproinsulin)	–AAAQHLCGSHLVDALYLVCGEKGFFYNPK-	GIVEQCCHKPCNIFDLQNYCN——
51	Salmon-3	1412252B(insulin B) +1412252A(insulin A)	–AAAQHLCGSHLVDALYLVCGEKGFFYTPK	G-IVEQCCHRPCNIFDLQNYCN——
52	Salmon-4	CAA31910(preproinsulin)	–AAAQHLCGSHLVDALYLVCGEKGFFYTPK-	GIVEQCCHKPCNIFDLQNYCN——
53	Salmon-5	CAA24983(preproinsulin)	–AAAQHLCGLHLVDALYLVCGEKGFFYTPK	G-IVEQCCHRPCNIFDLQNYCN——
54	Salmon-6	1303322A(insulin)	–AAAQHLCGSHLVDALYLVCGEKGFFYTPK	G-IVEQCCHRP———-
55	Butterflyfish	AAK28712(preproinsulin)	–ASSQHLCGSHLVDALYMVCGEKGFFYQPKT	GIVEQCCHHPCNIFDLQNYCN——
56	Cod	INCD(insulin)	-MAPPQHLCGSHLVDALYLVCGDRGFFYNPK-	GIVDQCCHRPCDIFDLQNYCN——
57	Leopard Frog	AAF87285(preproinsulin)	–FDNQYLCGSHLVEALYMVCGDRGFFYSPRS	GIVEQCCHNTCSLYDLENYCN——
58	African Clawed Toad-1	NP_001079351 (insulin precursor)	–LVNQHLCGSHLVEALYLVCGDRGFFYYPKV	GIVEQCCHSTCSLFQLESYCN——
59	African Clawed Toad-2	NP_001079350 (insulin precursor)	–LANQHLCGSHLVEALYLVCGDRGFFYYPKI	GIVEQCCHSTCSLFQLENYCN——
60	Nematode	AAC33275 (insulin-like peptide)	–DASIRLCGSRLTTTLLAVCRNQLCTGLTAF	GIATECCEKRCSFAYLKTFCCNQDDN

REFERENCES

[1] E. N. Baker, T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. M. C. Hodgkin, R. E. Hubbard, N. W. Isacs, C. D. Reynolds, K. Sakabe, N. Sakabe and N. M. Vijayan, *The structure 2 Zn pig insulin crystal at 1.5 A resolution*, Philos. Trans. R. Soc. London Ser. B 319: 369-456, 1988.
[2] J. Baussand and A. Carbone, *Inconsistent distances in substitution matrices can be avoided by properly handling hydrophobic residues*, Evol. Bioinform online, 4, 255-261, 2008.
[3] D. Burago, Y. Burago and S. Ivanov, *A course in Metric Geometry*, GSM, 33, American Mathematical Soceiety. 2001.
[4] J. M. Conlon, *Molecular Evolution of Insulin in Non-Mammalian Vertebrates*, Amer. Zool., 40: 200-212, 2000.
[5] R. H. M. Ebberink, A. B. Smit and J. Van Minnen, *The insulin family: evolution of structure and function in vertebrates and invertebrates*, Biol, Bull, 177: 176-182, 1989.
[6] F. M. Gregoire, N. Chomiki, D. Kachinskas and C. H. Warden, *Cloning and developmental regulation of a novel member of the insulin-like gene family in Caenorhabditis elegans*, Biochemical and Biophysical Research Communications; 249, 385 - 390, 1998.
[7] M. Kimura, *The neutral theory of molecular evolution*, Cambridge University Press, 1983.
[8] H. Koshiyama, *Explanation of the Insulin Paradox From the Evolutionary Point of View*, Jpn Clin Med. 2012; 3: 2124, 2012.
[9] C. Kristensen, T. Kjeldsen, F. C. Wiberg, L. Schaffer, M. Hach, S. Havelund, J. Bass, D. F. Steiner and A. S. Andersen, *Alanine scanning mutagenesis of insulin*, J. Biol. Chem, 272: 12978-12983, 1997.
[10] Y. Kunihiro and S. V. Sabau, *Quasi-metrics. Geometry of Sequence Comparison*, Proceedings of Asia Symposium on Engineering and Information, 2013, p. 186-196.
[11] H. P. A. Kunzi and V. Vajner, *Weighted quasi-metrics, in: Papers on General topology and Applic.*, Annals New York Acad. Sci. 728, 64-77, 1994.
[12] S. G. Mattews, *Partial metric topology*, in: Papers on General topology and Applic., Ninth Summer Conf. Slippery Rock, PA, Annals of the New York Acad. Sci. 728, 183-197, 1993.
[13] J. Pevnser, *Bioinformatics and Functional Genomics*, Second Edition, 2003.
[14] S. V. Sabau, K. Shibuya and H. Shimada, *Metric structures associated to Finsler metrics*, arXiv: 1305. 5880 [math.DG], 2013.
[15] A. Stojmirović and Y. Yu, *Geometric aspects of biological sequence comparison*, J. Comput. Biol., 16(4), 579-601, 2009.

Yuhei Kunihiro has graduated Tokai University, Department of Human Science and Informatics in 2011 and in present is graduate student in the Master Course at Tokai University, School of Science and Engineering, 5-1-1-1, Minami-Sawa, Minami-ku, Sapporo 005-8601, Japan e-mail: 2bsim001@mail.tokai-u.jp.

Sorin V. Sabau took his Ph. D. in Mathematics from Tokyo Metropolitan University and in present belongs to Tokai University, School of Science, Department of Mathematics, 5-1-1-1, Minami-Sawa, Minami-ku, Sapporo 005-8601, Japan e-mail: sorin@tspirit.tokai-u.jp.
His research encompasses Differential Geometry, Applied Mathematics and Bioinformatics.

Kazuhiro Shibuya took his Ph. D. in Mathematics from Hokkaido University and in present belongs Hiroshima University, Department of Mathematics, 1-7-1, Kagamiyama, Higashi-Hiroshima, 739-8511, Japan e-mail: shibuya@hiroshima-u.ac.jp.
His research encompasses Exterior Differential Systems, Differential geometry and applications.