Rigid geometry solves "curse of dimensionality" effects: an application to proteomics

Shun Adachi*

Center for Anatomical, Pathological and Forensic Medical Researches, Graduate School of Medicine, Kyoto University, Konoe-cho, Yoshida, Sakyo-ku, Kyoto, Kyoto 606-8501 Japan

*To whom correspondence should be addressed.

Abstract

Motivation: Quality of sample preservation at ultralow temperatures for a long term is not well studied. To improve our understandings, we need an evaluation strategy for analyzing protein degradation or metabolism at subfreezing temperatures. In this manuscript, we obtained LC/MS (liquid chromatography-mass spectrometry) data of calculated protein signal intensities in HEK-293 cells to monitor them.

Results: Our first trial for directly clustering the values has failed in proper arrangement of the sample clusters, most likely by the effects from "curse of dimensionality". By utilizing rigid geometry with *p*-adic (*I*-adic) metric, however, we could succeed in rearrange the sample clusters to meaningful orders. Thus we could eliminate "curse of dimensionality" from the data set. We discuss a possible interpretation for a group of protein signal as a quasiparticle Majorana fermion. It is possible that our calculation elucidates a characteristic value of a system in almost neutral logarithmic Boltzmann distribution of any type.

Contacts: f.peregrinusns@mbox.kyoto-inet.or.jp

1 Introduction

Even frozen, biological samples are said to be degraded during aging, and most frozen cell cultures are stored until they aged for two years. However, what actual happens in those samples are not well studied, so far as we know. There are a few reports that describe the existence of enzymatic activities in frozen cultures, such as lipase and peroxidase activities (e.g. Parducci and Fennema 1978; Voituron *et al.* 2006). However, we still do not know proteomic details of cells stored at subfreezing temperatures. For LC/MS (liquid chromatography-mass spectrometry), the only report we know dealing with cooled environments is the report for frogs whose environments mimicked the environments of winter (Kiss *et al.* 2011). This report lacks solid statistic analysis and it is not for subfreezing environment. Therefore we need solid proteomic data set from actual frozen cultures under long term storage at subfreezing environments to evaluate the potential degradation/metabolism.

To do this, first we need to set up an evaluation procedure that can well distinguish the samples from long term storage from the samples freshly prepared. Clustering analyses are popular approaches for the evaluation. Based on particular criteria that can evaluate similarity/dissimilarity, clustering analyses can observe meaningful groups in the data. The approaches are based on bottom-up calculation of the data, and there is no criterion outside of the system. However, there still remain problems such as how we define the groups and the selection of actual clustering methods. If the topological structure of the hierarchical tree or index numbers of clustering group are the same among all the different clustering methods, the output of the analyses is sound; however, the case is not always achieved: there might be some discrepancies and they cast doubt to the confidence of the results.

Mainly there are two types of clustering analysis: hierarchical clustering and non-hierarchical clustering. Hierarchical clustering can be calculable if there is a certain sort of distance/dissimilarity of the data point, and is able to join the data point based on close relationships among the point, until it can combine all the observed data set. Roughly speaking, it reduces multidimensional data to two dimensional data, with data labeling axis and clustering distance axis. The representative methods are: simple linkage, complete linkage, group average, weighted average, centroid, median, Ward's method. If we set dissimilarity of i, j, k as C_i, C_j, C_k,

$$d\left(C_i \cup C_j, C_k\right) = \alpha_i d(C_i, C_k) + \alpha_j d\left(C_j, C_k\right) + \beta d\left(C_i, C_j\right) + \gamma\left|d(C_i, C_k) - d\left(C_j, C_k\right)\right|,$$

and the values of α, β and □ are described in Table 1.

Table 1. Parameters of hierarchical clustering methods and their evaluation.

method	α_i	α_j	β	□	dissimilarity	monotony	metric
single	1/2	1/2	0	-1/2	no restriction	T	reduction
complete	1/2	1/2	0	1/2	no restriction	T	expansion
group average	$n_i/(n_i + n_j)$	$n_j/(n_i + n_j)$	0	0	no restriction	T	conserved
weighted average	1/2	1/2	0	0	no restriction	T	conserved
centroid	$n_i/(n_i + n_j)$	$n_j/(n_i + n_j)$	$-n_i n_j/(n_i + n_j)^2$	0	*E2*	F	conserved & reduction
median	1/2	1/2	-1/4	0	*E2*	F	conserved & reduction
Ward	$(n_i + n_k)/(n_i + n_j + n_k)$	$(n_j + n_k)/(n_i + n_j + n_k)$	$-n_k/(n_i + n_j + n_k)$	0	*E2'*	T	conserved & expansion

E2, *E2'*, T, F are square Euclid distance, a half of square Euclid distance, true and false, respectively. It is easily recognizable that β, □ values are correction values based on a triangle of *i*, *j*, *k*. Metric expansion/reduction mean renewal of ongoing clustering by farther distance of each data length/vice versa. Monotony means clustering of lengths at each calculation step is monotonically increasing; which is not true in centroid and median methods. Depending on the situation, either T or F case is favored.

Non-hierarchical clustering, for example *k*-means method, is an optimization method based on portioning of groups and classification. First of all, we have to set the number of groups, *k*, among the data set. As dissimilarity, we use square Euclid distance. After that, we set initial grouping and score each group, and put samples one by one. We can select the score of lower case and repeat the process. In the sense that we should select the number of groups it is top-down approach, but from other aspects it is bottom-up. All these eight methods are easy to be equipped in computer and very frequently used, compared to other complex methodologies.

One problem for the analyses is, due to high dimensionality (more than 1000) of the samples, there is “curse of dimensionality” effect and the variances among samples become large and sparse, resulting in meaningless output of the clustering analysis (e.g. Ronan *et al*. 2016). A solution of this procedure is utilizing machine learning method for evaluation. However, the high dimensionality results in very large or sometimes incalculable value of Akaike information criterion, which exhibits a doubt to the solution appeared.

In this manuscript, we show you another mathematical solution for pretreatment of clustering analyses, based on rigid geometry. The most important part of clustering is what type of metric we use in the analyses, and we would like to say *p*-adic metric (of prime ideal) based on rigid geometry is very favorable to discriminate control samples and samples endured long-term storage. As a definition of metric, (i) it satisfies separation axiom (not necessarily non-negative); (ii) the identity of indiscernibles; (iii) it satisfies symmetry; (iv) it satisfies triangle inequality. As examples of metric, there are absolute distance, Chebyshev distance, Euclid distance, average Euclid distance, square Euclid distance, Minkowski distance, correlation efficiency, cosine efficiency etc. The selection of metric gives significant difference in output of the calculation (e.g. Ronan *et al*. 2016). The common principle for modern geometry is that it converts the observed values to either nilpotent state for convergence/divergence or -1 state for oscillation, to handle the system easier than the original system. Rigid geometry, which is most famous among similar mathematical field, based on complete non-archimedean field and has a history from 1962 introduced by John Tate (Tate 1971; Fujiwara and Kato 2006; Kato 2011), utilizing *p*-adic elliptic curve to solve the situation. Non-archimedean valuation of the system enables converged values in global, but locally freed values from high dimensionality. The topology is called *G*(Grothendieck)-topology. Probably we would be able to demonstrate the first successful example of applying rigid geometry to biological data set, exhibiting freedom from "curse of dimensionality" effects that unifies the output topology of clustering without investing difficulty in interpreting the clustering results.

2 Methods

2.1 Cell culture

Human HEK-293 cell line originated from embryonic kidney was purchased from RIKEN (Japan). The original cultures were frozen on either March the 18th 2013 (3 yrs sample) or March the 5th 2014 (2 yrs sample), and used in experiments between February and June 2016. The strain was cultured in MEM (Modified Eagle's Medium) + 10% FBS (Fatal Bovine Serum) + 0.1 mM NEAA (Non-Essential Amino Acid) at 37°C with 5% CO_2. Subculturing was performed in 0.25% trypsin, and original cells from RIKEN before our experiments, were frozen obeying the standard protocol of RIKEN: in culture medium with 10% DMSO (dimethyl sulfoxide), they were cooled until reaching 4°C at -2°C/min, stayed for 10 min, frozen until reaching -30°C at -1°C/min, stayed for 10 min, and cooled until reaching -80°C at -5°C/min and stayed overnight. The next day, they were transferred to liquid nitrogen storage. Freezing conditions for actual control experiments in our study are described in Results section. In brief, 'fresh' means fresh samples immediately underwent protein extraction processes. '1 h' means they were harvested and stayed for 1 h at -80°C with freezing medium. 'o/n-o/n' means they were harvested and stayed overnight at -80°C with freezing medium, then transferred to liquid nitrogen storage overnight.

2.2 Protein extraction, alkylation and digestion

Proteins of HEK-293 were extracted by the standard protocol of RIPA Buffer (nacalai tesque, Inc., Kyoto, Japan). In brief, ~10^6 of harvested cells were washed in Krebs-Ringer-Buffer (KRB; 154 mM NaCl, 5.6 mM KCl, 5.5 mM glucose, 20.1 mM HEPES (pH 7.4), 25 mM $NaHCO_3$) once. They were resuspended in 30 µl of RIPA Buffer, taking in and out through 21G needles for destruction and incubated on ice for 1 h. Then they were centrifuged at 10,000 *g* for 10 min at 4°C, followed by collection of supernatants, quantified the amounts of proteins by Micro BCA Protein Assay Kit (ThermoFisher SCIENTIFIC, Inc., Waltham, U.S.A.) and continued to the processes of XL-Tryp Kit Direct Digestion (APRO SCIENCE, INC., Naruto, Japan). The samples were solidified in acrylamide gels, washed twice in ultrapure water, washed again three times in dehydration solution, and dried. Then the samples were continued to the processes in In-Gel R-CAM Kit (APRO SCIENCE, INC., Naruto, Japan). The samples were reducted for 2 h at 37°C, alkylated for 30 min at room temperature, washed five times with ultrapure water and twice with destaining solution, then dried. The resultant samples were trypsinized overnight at 35°C. The next day, digested peptides dissolved were collected by ZipTipC18 (MERCK MILLIPORE, CORP., Billerica, U.S.A.). The tips were dampened with acetonitrile twice and equilibrated twice by 0.1% trifluoroacetic acid. The peptides were collected by ~20 cycles of aspiration and dispensing, washed with 0.1% trifluoroacetic acid twice and eluted by 0.1% trifluoroacetic acid /50% acetonitrile with aspiration and dispensing five times × three tips followed by vacuumed drying up. The finalized samples were stored at -20°C. Before performing LC/MS, they were resuspended in 0.1% formic acid, and the amounts were quantified by Pierce Quantitative Colorimetric Peptide Assay (ThermoFisher SCIENTIFIC, Inc., Waltham, U.S.A.).

2.3 LC/MS

LC/MS was performed by Medical Research Support Center, Graduate School of Medicine, Kyoto University with Quadrupole-Time of flight [Q-Tof] type mass spectrometer TripleTOF 5600 (AB SCIEX Pte., Ltd., Concord, Canada). The procedures obeyed their standard protocols. The loading amounts for each sample were 1 µg. The obtained quantitative data for identified proteins as Unused information were extracted by ProteinPilot 4.5.0.0 software (AB SCIEX Pte., Ltd., Concord, Canada).

2.4 Clustering analyses and machine learning of the pattern

Hierarchical clustering analyses were performed by standard hclust function in R 3.2.3 (https://cran.r-project.org). The actual hierarchical methods used were: single linkage; complete linkage; group average; weighted average; centroid; median; Ward’s method. *k*-means method was performed by standard kmeans function in R 3.2.3. It was calculated based on all eleven samples. As machine learning program, 11-1-1 hierarchical neural network analysis was performed in R 3.2.3 with a package nnet and factors cl (the number of raw) were calculated as a characterization index of pattern. For calculation, we only used Unused values appeared in all the eleven samples to avoid distortion of the calculation come from failures in identification within LC/MS, not from significantly small signal values (N = 800). The actual data of Unused values for calculation are shown in Table S1.

2.5 Utilizing a *p*-adic (*I*-adic) metric embedded on rigid geometry

Now we set an analogy (a grounding metaphor) of biological data space (as base) and mathematical space (as target). We will not get into details of the opposite direction of analogy, as we still do not understand how the target space behaves in details mathematically. We will think a projection from the base to the target, and also utilize theories for formal schemes in analyses of projected data (linking metaphors). The improvement of mathematical metrics data of Unused values was directed by ideas in rigid geometry as follows. Please also refer Adachi (2016). In brief, the data from each sample were first arranged in their ranks k of Unused values N_k, approximated by logarithmic approximation:

$$N_k = a - b \ln k.$$

The actual R^2 values were approximately 0.84-0.95. Then

$$\mathfrak{R}(s) = \frac{\ln \frac{a}{N_k}}{\ln k}$$

was calculated as a deviation index from logarithmic distribution. After that,

$$|D| = e^{\frac{\mathfrak{R}(s)}{b}}$$

were calculated as an index of absolute fitness values for ln(protein density) predicted from logarithmic Boltzmann distribution of the protein signals. Then, average signal values of the sample $E(N)$ was calculated and an expected overall cooperative fitness of each protein for their promising future

$$p = |D|^{E(N)}$$

was calculated as p values of a p-adic (or prime ideal I-adic) metric of prime ideals. A Euclidean metric of a complex s = Re(s) + $p\surd$-1, $\sqrt{(\mathrm{Re}(s_2) - \mathrm{Re}(s_1))^2 + (p_2 - p_1)^2}$ obviously satisfies (i) separation axiom; (ii) the identity of indiscernibles; (iii) symmetry; (iv) triangle inequality. s values also fulfill the requirement as a parameter of a high-dimensional theta function that converged absolutely and uniformly on complex three-dimensional compact subset (Neukirch 1999; Adachi 2016). To understand this, we would relate the theta function to the upper half plane $\mathbb{H} \ni s$. Setting complex

$\frac{1}{\zeta k^s} \in \mathbb{C}$, $\mathbb{R} = [\prod \mathbb{C}]^+ = \{s \in \mathbb{C} \mid s = \bar{s}\}$, and Hecke ring $\mathbb{R}$ is the Minkowski space. Then dual space of the Hecke ring $\mathbb{R}_+^* \ni p$ and natural logarithm of it, $\{\mathbb{R}_\pm | \mathbb{R}_\pm \ni E(N)\ln|D|\} \subseteq \{\mathbb{R}_b | \mathbb{R}_b \ni b\ln|D|\}$. We can successfully define $\{s \in \mathbb{H} \mid \mathbb{H} = \mathbb{R}_\pm + \sqrt{-1}\mathbb{R}_+^*\}$.

$$\mathbb{H} \subseteq \mathbb{C} \supseteq \mathbb{R} \supseteq \mathbb{R}_\pm \supseteq \mathbb{R}_+^*$$

is the form we need for constituting theta function converging absolutely and uniformly on every compact subset $\mathbb{R} \times \mathbb{R} \times \mathbb{H}$ (Neukirch 1999). We can say *s*, and especially *p*, have quasi-compactnesses and quasi-separations.

Finally, a flag manifold,

$$v = \frac{\ln N_k}{\ln p}$$

was calculated and a set of *v* is a coherent rigid analytic space with a coherent formal scheme $\{\mathbb{S}_F | \mathbb{S}_F \ni \ln N_k\}$ (it has to be an adequate formal scheme but not necessary to be Noetherian scheme and thus we enable to include non-Noetherian schemes such as irreversible time-asymmetric model) divided by a *p*-adic blow-up ln*p* with quasi-compactness and quasi-separation (Fujiwara and Kato 2006; Kato 2011). *If we regard* $\ln N_k$ *is on Tate Algebra (this is a single assumption required for this work)*, a quotient by an ideal ln*p* is an isomorphic to *k*-Banach Algebra, which is an affinoid algebra with quasi-compactness and quasi-separation. *v* (in original non-archimedean it should be –*v*, however, the opposite sign does not make difference in further discussion as *v* metric) becomes affinoid in locally closed immersions among the affinoid varieties when the projections are bijective, neglecting the case *p* = 1 (Gerritzen and Grauert 1969; Temkin 2005). From arithmetic calculations based on -*v* N_k space *v* is on an ultrametirc space, but now we calculate *v* from N_k and only think of *v* on Euclidean metric space. A Euclidean metric of *v*, $\sqrt{(v_2 - v_1)^2}$ obviously satisfies (i) separation axiom; (ii) the identity of indiscernibles; (iii) symmetry; (iv) triangle inequality. We do not use the top *k* = 1 proteins for analyses due to the impossibility of the calculation by 1/ln*k* reaching infinity.

For the proof that *v* obeys rigid geometry, first, we show you Schottky-type uniformization of the elliptic curve on the complex. Let periodic lattice Λ, expected *N* to be *E*(*N*) and a normalization factor of *P* as $N_k/E(N) = PD^{Nk}$;

$$\Lambda = 2\pi i(\mathbb{Z} + \tau\mathbb{Z}),$$

$$(\tau \in \mathbb{H} = \{z \in \mathbb{C} | \mathfrak{J}(z) > 0\}),$$

$$\ln p^{v^{\square}} = \ln N_{k_{\square}} \in \mathbb{C}.$$

$$\mathbb{C} \xrightarrow{\square exp} \mathbb{C}^{\times} \ni N_k = p^v,$$

$$\mathbb{C}/\Lambda = \mathbb{C}^{\times}/q^{\mathbb{Z}} = \left\{ \mathbb{S} \mid \mathbb{S} \ni PE(N)D \bmod N_k = \frac{PD^{N_k}}{\frac{N_k}{E(N)}} \right\}.$$

Infinite dimensional covering of the last sentence is Schottky-type uniformization (c.f. Fujiwara and Kato 2006). Note that $|q|_p < 1$. This is identical to geometric type of prime number theorem $PE(N)\pi^G(x) \sim PE(N)e^x/x$, where $\pi^G(x)$ is a prime counting function of the value $x = N_k$. On an Weierstrass elliptic curve E: $y^2 = 4x^3 - g_2x - g_3$, Tate curve can be realizable only if

$$|j(E)|_p = \left| 1728 \frac{g_2^3}{g_2^3 - 27g_3^2} \right| > 1,$$

which means assuming $g_2 << 1$ and $|g_3| >> 1$ would result in collapse of a Tate system. That is, if the relative effects from outside world on x variant is too large, the interacting efficiency of the x constitutes y^2 would lose the identity of the system, and this limit the size of system.

Then, let M be a differentiable manifold; $\Omega^0(M)$ be a space of smooth function on a rigid analytic space M, $\Omega^i(M)$ be a space of i-th differential form. d^i: $\Omega^i(M)$ -> $\Omega^{i+1}(M)$ represents exterior derivative and elements of Ker d^i and Im d^i are closed form and exact form, respectively. d^{i+1} $d^i = 0$ and

$$0 \to \Omega^0(M) \to \Omega^1(M) \to \Omega^2(M) \to \Omega^3(M) \to \cdots$$

is a crystalline complex, cohomological to crystalline cohomology (Grothendieck, 1966; 1968). That is,

$$H_{dR}^i(M) = \mathrm{Ker}\ d_i/\mathrm{Im}\ d_{i-1}$$

is an i-th crystalline cohomology group. Therefore $H^i_{dR} = 0$ and that ‘any i-th closed form is exact form’ are equivalent. We can take a set of rigid analytic space, modular N_k as Ω. Please note that setting p as an element of Coxeter group, an identity element of p corresponds to an identity element of Hecke ring. $d = p$ is thus proper. Furthermore, $i = v$ is smooth when $p \neq 1$. Since exterior derivative of p is 0 and obviously p is exact form unless $v = 0$, $H^i_{dR} = 0$. $v = \ln N(t)/\ln p$ (t is time) is obviously on unit polydiscs of rigid analytic space, rendering locally ringed G-topologized space with a sheaf of non-archimedean field, which ensures a covering by open subspaces isomorphic to affinoids. Shifting –v (non-archimedean for $1/N_k = p^{-v}$) to v metric (non-archimedean for $N_k = p^v$) does not change this property, considering $1/N_k$ space as basis. In other words, $\ln N_k$ is related to the kernel of present signal space and $\ln p$ is related to the kernel of potential signal space, which is the image of past signal space. The division of them, v is the image of potential signal space, which reflects the physiological situation of the system adapted to expecting environments without any noise of current system. Overall, the system described here has a rigid cohomology (Kedlaya 2009). Considering $N_k = PD^{Nk}$ and $P = 1/(D^a\zeta(s))$ in this case (Adachi 2016),

$$v = \frac{N_k \ln D - a \ln D - \ln \zeta(s)}{E(N) \ln|D|}$$

and overconvergence of the v values is thus achieved due to cancelling out of high dimensionality in N, a together with topological characteristics (G-topology) of v on quasi-compactness and quasi-separation as mentioned before.

3 Results

3.1 Direct analyses of Unused values in LC/MS resulted in non-proper clustering of samples either by hierarchical or *k*-means clustering method, but showed obscure patterning by machine learning

First, we extracted proteins from HEK-293 cells that underwent different freezing conditions. As a control, we collected ‘fresh’ samples from culturing cells, ‘1 h’ samples from the cells frozen for 1 h at -80°C, and ‘o/n-o/n’ samples from the cells frozen overnight at -80°C and subsequently transferred to liquid nitrogen overnight. For the samples of interest we used the sample preserved in liquid nitrogen for 2 or 3 years. See Methods for more detail. After performing LC/MS, we extracted information of Unused values and performed clustering analyses of various hierarchical clustering methods and also *k*-means methods. As results, the fluctuation based on different experiments affect the data to significant extents, such that we could not obtain any meaningful clustering by these methods as shown in Fig. 1. The control samples and the samples with long-term storage were mixed up. By machine learning based on neural network, however, exhibited clustering of cl values on both control samples and storage samples. The samples from 2-years-preservation varied from the control values, and an only 3-years sample occupied between those clusters (Fig. 1). These results suggest that there might be “curse of dimensionality” effects, which based on significantly varied values of each data point in high dimensionality that disturbs convergence of the output values (e.g. Ronan *et al.* 2016). To confirm this idea, the number of unknown parameters in neural network is 1630. The image of these ideas is described in Fig. 2A. Obviously actual structures of geometric space are important for the resultant output of the calculation (Ronan *et al.* 2016). In this first analyses, we used simple Unused values as dissimilarity, partly in square Euclidean distance.

3.2 A *p*-adic metric based on rigid geometry eliminated "curse of dimensionality" effects on LC/MS data

To avoid the pitfalls described above, the first choice as a solution is to design more proper metric for calculations. If we set a proper metric for calculation based on geometry, which enables nilpotent for convergence/divergence of values and converging to the value of -1 as oscillation, we can extract more overconverged output from the observed data set to discriminate the characteristics observed. One of the popular methods for this trial is rigid geometry. Non-archimedean valuation field in the geometry is easy to converge compared to Archimedean real filed or complex field, with *p*-adic (*I*-adic) metric including a subring of norm < |1|. The geometry globally converges the values, but the values are locally free, enabling freedom from the restriction by "curse of dimensionality". The example image to utilize this idea by quotient is described in Fig. 2B (e.g. Cornelissen and Kato 2005). Consider icosahedron with 12 vertices in blue color, 20 barycenters (the center of the triangle with 20 faces) in green color, 30 edges with 30 midpoints in red color. Projecting the icosahedron from its center to a sphere maps tessellation of the sphere by 120 triangles as shown in Fig. 2B left. The angles are $\pi/2$ for red, $\pi/3$ for green, and $\pi/5$ for blue. A generator is:

$$I = \langle \begin{pmatrix} \zeta & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \zeta + \zeta^{-1} & 1 \\ 1 & -(\zeta + \zeta^{-1}) \end{pmatrix} \rangle, \zeta = e^{2\pi i/5}$$

The icosahedron has 6 cyclic subgroups of order 5, 10 cyclic subgroups of order 3, and 15 cyclic subgroups of order 2. The quotient of this Riemann sphere by the group *I* is shown in Fig. 2B right. 2, 3, 5 correspond to midpoints of edges, barycenters of faces, and vertices, respectively. The complexity of the system is much more simplified.

Now we defined a *p*-adic (*I*-adic) metric based on rigid geometry as in Methods section as a pretreatment of data before clustering/machine learning, and obtained the results as Fig. 3. Obviously control samples and samples of long-term storage clustered separately in any type of proposed methods, suggesting freedom from "curse of dimensionality". Although the means of variances in the original method and the rigid method do not represent the situation (60±10 and 6000±8000 for 95% confidential, respectively). The data from the rigid method have 10 outliers (See Fig. 4 for the skewness) that have larger values than Euclidian values of the same ranks. When samples of top 10 variances are excluded, the means of variances become 44±5 and 14±3, respectively, with $p = 5 \times 10^{-20}$ for *t*-test, indicating release from "curse of dimensionality". As a control, machine learning by neural network showed the same tendency as the previous section, with the number of unknown parameter value as 1630.

Interestingly, the distribution of v values can be approximated by a power function with an absolute value of multiplier as ~3/2 (Fig. 4). If we set the multiplier as Hurwitz-Kronecker class number $H(d_H)$ (Adachi 2016), $d_H = 16$ (Zagier 2000) and the dimension corresponds to a non-anomalous weight of 16 almost without internal interactions in Tachikawa and Yonekura (2016) and Witten (2016), on ν bands of Majorana fermion with 3 + 1 or 2 + 1 (3 spatial dimensional N_k and time dimension, or the parameter s of 2 real dimension and time-dimension) spacetime. For Majorana fermion, the particle and antiparticle are the same and in condensed matter physics, it can be regarded as quasiparticle. That is, if a group of a protein behaves as almost a single type of population regarding the synthetic fitness of it, it can be regarded as Majorana fermion.

4 Discussion

Utilizing *p*-adic rigid geometry, we seem to succeed in eliminating the "curse of dimensionality" effects from significantly diverged sample data, at least in LC/MS data set of HEK-293. Basically if the sum of the number of dimensions ($\sum n_d$) exceed the original number of model dimensions (in our case it should be $n = 1630$, according to neural network), the values could be converged assuming $\sum n_d - n$ number of traces becomes nilpotent (Weyl 1953). 800 x 11 = 8800 > 1630 and the observed convergence of v is expected beyond underdetermined system. To support this idea, clustering of f-1, f-2 and 2y2, which were mis-clustered in Fig. 1, could be clustered well in v metrics in all the methods (Fig. 5), with 800 x 3 = 2400 > 1606. At least this allows us to evaluate whether the samples are from nearly fresh materials or underwent significant lengths of storage at low temperatures. The success is entirely based on an algebraic, analytic and topological geometric analysis based on rigid geometry. So far as we know, this is the first work that applies ‘rigid geometry’, as the term developing in mathematical fields since 1962, to biological studies. The interesting point is that this methodology can be applicable to any type of almost neutral logarithmic Boltzmann-type distribution in any type of systems interested. The agreement of results in both a supervised machine learning and several unsupervised clustering analyses demonstrates the power of this methodology. Even in biology, we can apply similar approach from protein society inside cells in this study to community dynamics in microbes (Adachi 2016). Application to other research field such as chemistry, physics, astronomy and earth science, is promising if we can successfully introduce ‘fitness’ idea to the fields.

d_H = 16 case can be interpreted as (2 dimensional s × 4 usual spacetime dimensions) × 2 (interaction of the two particles) = 16. In other word 2^2^2 (three 2s with multipliers) = 16. This model neglects fluctuation to the other dimensions that results in 24 dimensions (See also Adachi 2016). In this case,

$$v = \frac{\ln N_k}{\ln p} = \frac{\psi_\alpha}{\psi^{\dot{\alpha}}} = \Psi$$

is a four-component Majorana fermion with weak coupling (Tachikawa and Yonekura 2016).

5 **Conclusions**

We have succeeded in the release from “curse of dimensionality” of observed difference among the samples of long term storages and control samples with LC/MS data in HEK-293 cells. The success was entirely based on topological characteristics of p-adic metric on rigid geometry. It may have a potential to calculate a characteristic value of a system with almost neutral logarithmic Boltzmann distribution of any type.

Acknowledgements

We thank Prof. Tatsuaki Tsuruyma for providing us facilities of experiments and funding. We also thank for performing LC/MS by Medical Research Support Center, Graduate School of Medicine, Kyoto University. Additional support was provided by the Center for Anatomical, Pathological and Forensic Medical Researches, Graduate School of Medicine, Kyoto University, Japan. We are grateful for these laboratorial and financial supports.

Funding

This work was supported by a Center of Innovation (COI) program from Japan Science and Technology Agency, regarding Prof. Tatsuaki Tsuruyama of Graduate School of Medicine, Kyoto University.

References

Adachi, S. (2016) Discrimination of domination mode and chaotic mode in species. arXiv:1603.00959v5 [q-bio.PE].

Cornelissen, G. and Kato, F. (2005) The *p*-adic icosahedron. *Notices Amer*. *Math. Soc.*, **52**, 720-727.

Fujiwara, K. and Kato, F. (2006) Rigid geometry and applications. *Advanced Studies in Pure Mathematics*, **45**, 325-384.

Gerritzen, L. and Grauert H. (1969) Die Azyklizität der affinoiden überdeckungen, in: Spencer, D.C. and Iyanaga S. eds. Global Analysis (Papers in Honor of Kodaira, K.). Univ. Tokyo Press, Tokyo, pp. 159-184.

Grothendieck, A. (1966) On the de Rham cohomology of algebraic varieties. *Institut des Hautes Études Scientifiques. Publications Mathématiques*, **29**, 95-103 (Letter to Atiyah, 14 Oct 1963).

Grothendieck, A. (1968) Crystals and the de Rham cohomology of schemes, in: Giraud, J. *et al.* eds. Dix Exposés sur la Cohomologie des Schémas. *Advanced Studies in Pure Mathematics*, **3**, North-Holland Publishing Co., Amsterdam, pp. 306–358.

Kato, F. (2011) Topological rings in rigid geometry, in: Cluckers, R. *et al*. eds. Motivic Integration and Its Interactions with Model Theory and Non-Archimedean Geometry, Vol. I. *London Math. Soc. Lecture Note Ser.*, **383**, Cambridge Univ. Press, Cambridge, pp. 103-144.

Kedlaya, K.S. (2009) *p*-adic cohomology, in: Abramovich, D. *et al*. eds. Algebraic geometry---Seattle 2005, Part 2. *Proc*. *Sympos*. *Pure Math.*, **80**, Amer. Math. Soc., Providence, pp. 667–684.

Kiss, A.J. *et al.* (2011) Costanzo, Seasonal variation in the hepatoproteome of the dehydration and freeze-tolerant wood frog, *Rana sylvatica*. *Int. J. Mol. Sci*., **12**, 8406-8414.

Neukirch, J. (1999) Algebraic Number Theory. Springer Verlag, Berlin-Heidelberg-New York.

Parducci, L.G. and Fennema, O. (1978) Rate and extent of enzymatic lipolysis at subfreezing temperatures. *Cryobiology*, **15**, 199-204.

Ronan, T. *et al.* (2016) Avoiding common pitfalls when clustering biological data. *Sci. Signal.*, **9 re6**, 1-12.

Tachikawa, Y. and Yonekura, K. (2016) Gauge interactions and topological phases of matter. arXiv:1604.06184v2 [hep-th].

Tate, J. (1971) [1962] Rigid analytic spaces. *Inventiones Mathematicae*, **12**, 257-289.

Temkin, M. (2005) A new proof of the Gerritzen-Grauert theorem. *Math. Annal.*, **333**, 261-269.

Voituron., Y. *et al.* (2006) Oxidative DNA damage and antioxidant defenses in the European common lizard (*Lacerta vivipara*) in supercooled and frozen states. *Cryobiology*, **52**, 74-82.

Weyl, H. (1953) The Classical Groups: Their Invariants and Representations, second revised ed. Princeton Univ. Press, Princeton.

Witten, E. (2016) The “parity” anomaly on an unorientable manifold. arXiv:1605.02391v2 [hep-th].

Zagier, D. (2000) Traces of singular moduli. SIS-2004-265, cds.cern.ch/record/738436/files/sis-2004-265.ps Accessed 21 Oct 2015.

Legends to Figures

Fig. 1. Clustering of Unused value sets of each protein in LC/MS (*N* = 800). f-1, 2, 3; freshly prepared 'fresh' samples in experiments 1, 2, 3, respectively. 1h1, 1h2; samples frozen at -80°C for 1 h ('1h') in experiments 1, 2, respectively. o/1, o/2; samples stayed at -80°C o/n and then in liquid nitrogen storage o/n ('o/n-o/n') in experiments 1, 2, respectively. 2y1, 2, 3; samples preserved in liquid nitrogen storage of RIKEN for approximately 2 years in experiments 1, 2, 3, respectively. 3y; a sample preserved in liquid nitrogen storage of RIKEN for approximately 3 years. The numbers in *k*-means method are the index numbers of classified groups. The numbers in neural network are factors cl values, which represents one-dimensional characteristics of the systems. Please also read Methods.

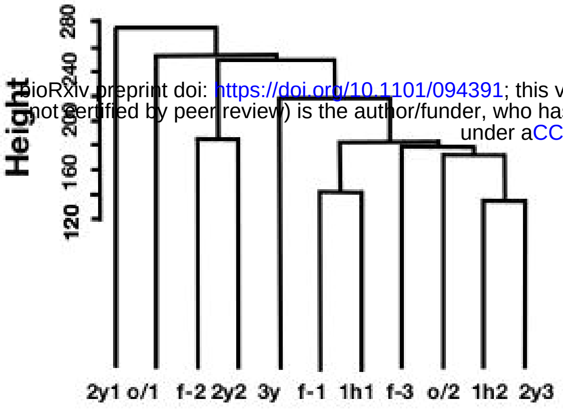
Fig. 2. (A) "Curse of dimensionality" effects, with sparser geometric distribution of data points. See also Ronan *et al.* (2016). **(B)** Example of geometric conversion to a simpler system: quotient of icosahedral tessellation on a Riemann sphere by *I*. See also Cornelissen and Kato (2005).

Fig. 3. Clustering of newly invented *v* value sets of each protein in LC/MS (*N* = 800). f-1, 2, 3; freshly prepared 'fresh' samples in experiments 1, 2, 3, respectively. 1h1, 1h2; samples frozen at -80°C for 1 h ('1h') in experiments 1, 2, respectively. o/1, o/2; samples stayed at -80°C o/n and then in liquid nitrogen storage o/n ('o/n-o/n') in experiments 1, 2, respectively. 2y1, 2, 3; samples preserved in liquid nitrogen storage of RIKEN for approximately 2 years in experiments 1, 2, 3, respectively. 3y; a sample preserved in liquid nitrogen storage of RIKEN for approximately 3 years. The numbers in *k*-means method are the index numbers of classified groups. The numbers in neural network are factors cl values, which represents one-dimensional characteristics of the systems. Please also read Methods.

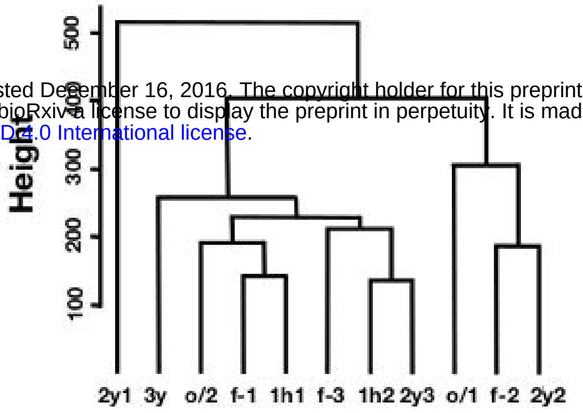
Fig. 4. Ranked variance distributions of Unused values and *v* values of proteins used for calculations (*N* = 800). x-axis; the rank of values. y-axis; the actual variance values.

Fig. 5. Clustering of newly invented *v* value sets of each protein in LC/MS (*N* = 800). f-1, 2; freshly prepared 'fresh' samples in experiments 1, 2 respectively. 2y2; a sample preserved in liquid nitrogen storage of RIKEN for approximately 2 years in the experiment 2. The numbers in *k*-means method are the index numbers of classified groups. The numbers in neural network are factors cl values, which represents one-dimensional characteristics of the systems. Please also read Methods.

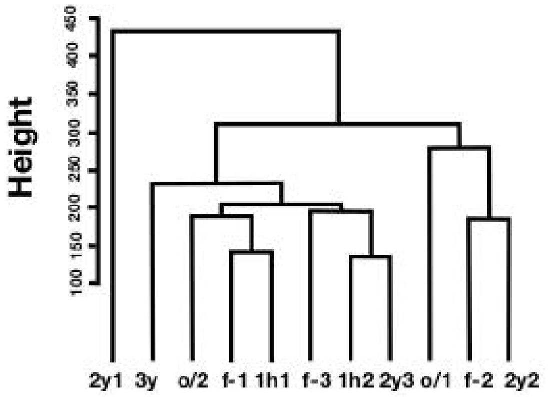
Table S1. The table of Unused values for identified proteins. Please also see the legend of Fig. 1.



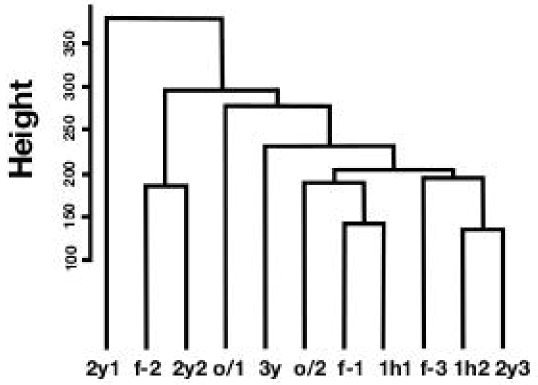
Single Linkage



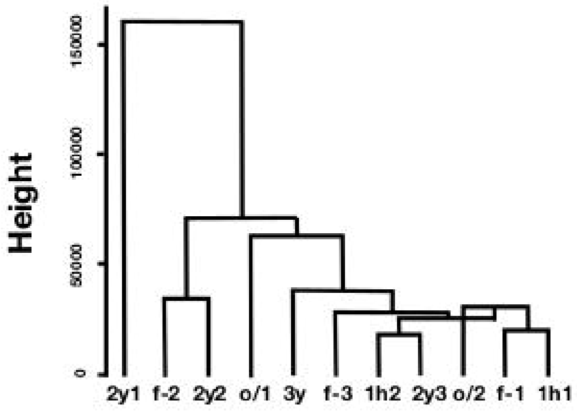
Complete Linkage



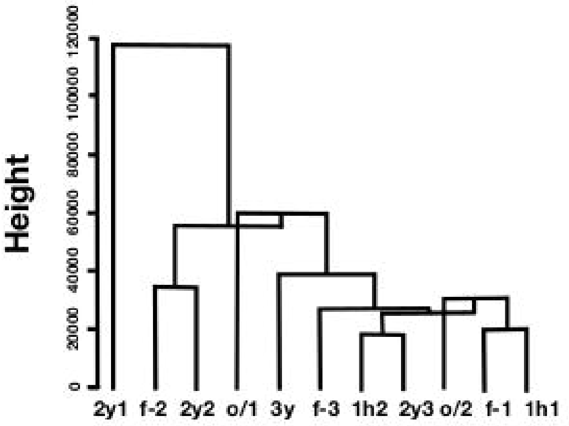
Group Average



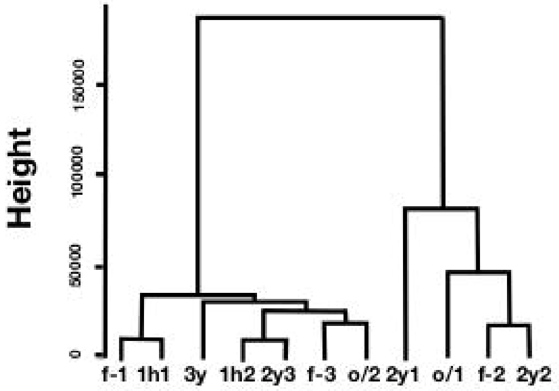
Weighted Average



Centroid



Median



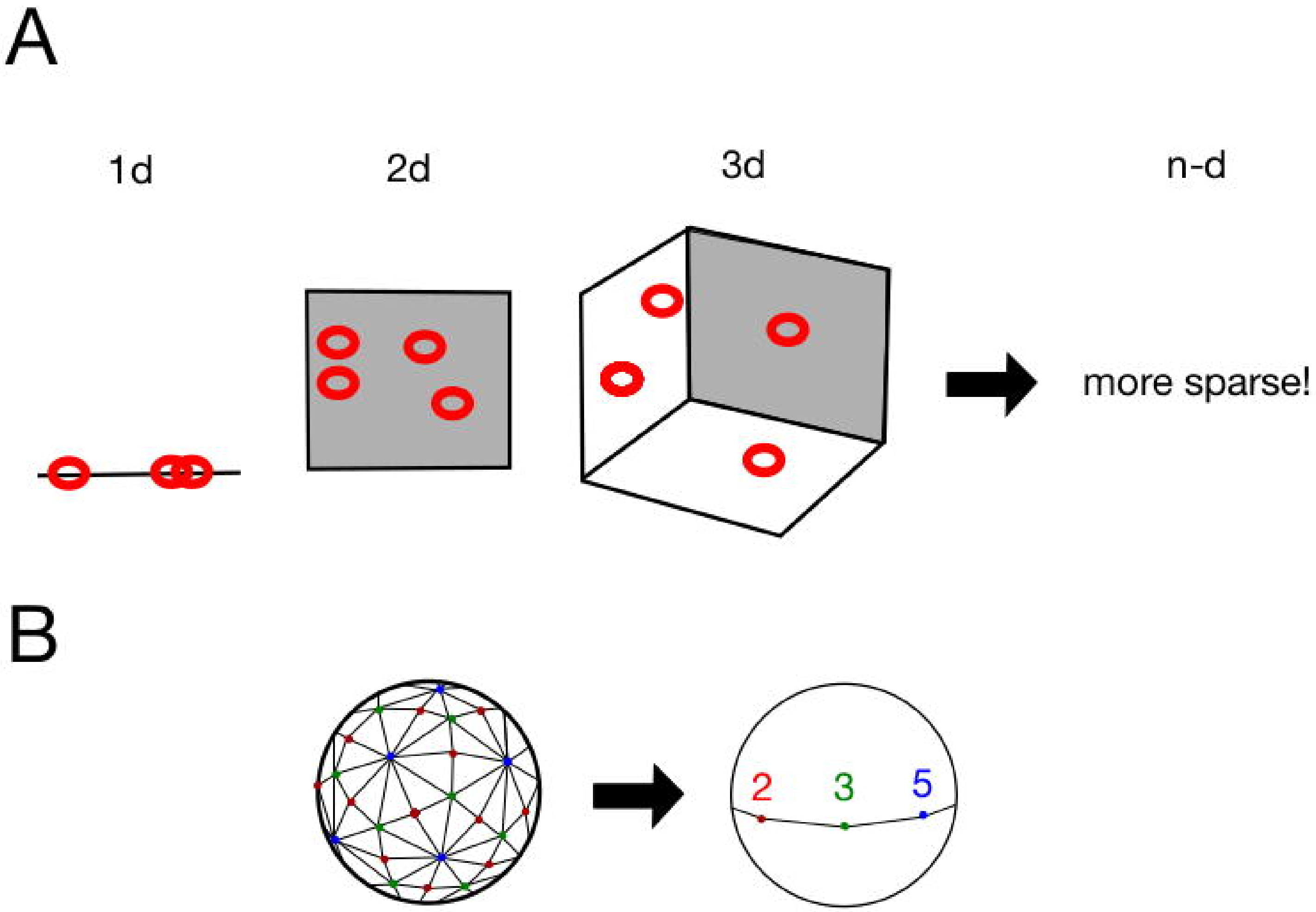
Ward's method

k = 5 means

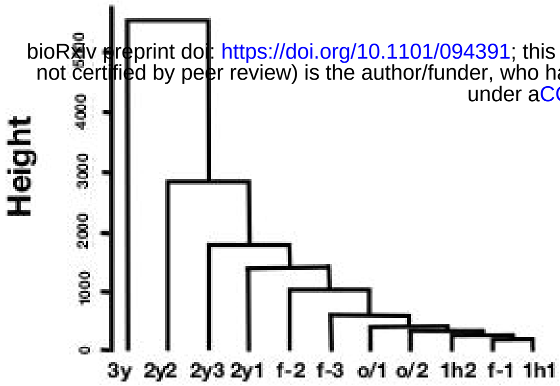
f-1	f-2	f-3	1h1	1h2	o/1	o/2	2y1	2y2	2y3	3y
3	5	2	3	4	4	3	1	5	4	2

Neural Network

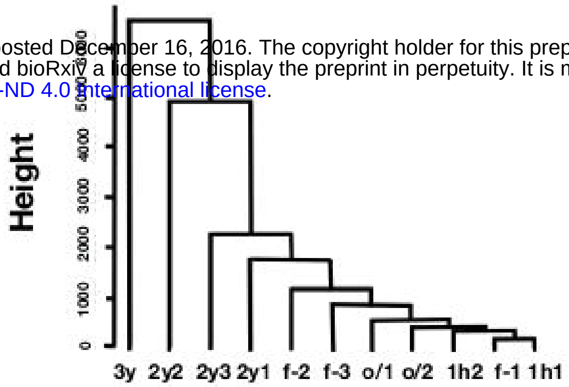
f-1	f-2	f-3	1h1	1h2	o/1	o/2	2y1	2y2	2y3	3y
17.35	16.68	18.33	23.4	15.21	13.24	19.4	0.89	6.27	8.88	13.03



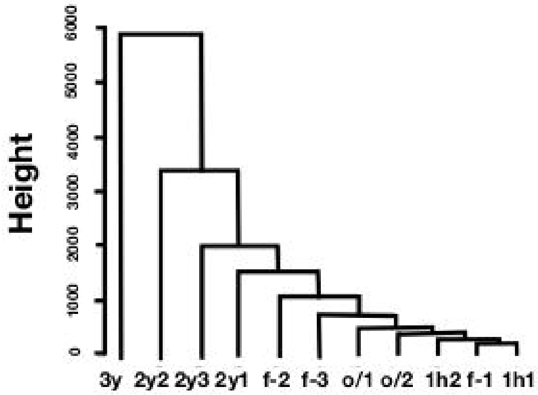
A
1d
2d
3d
n-d
more sparse!
B
2
3
5



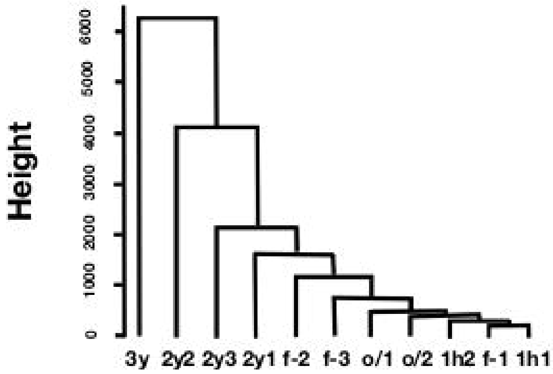
Single Linkage



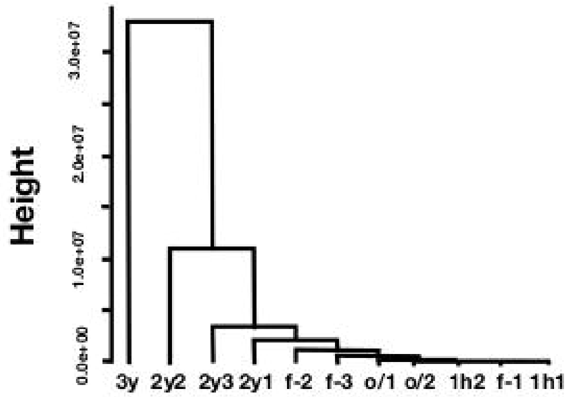
Complete Linkage



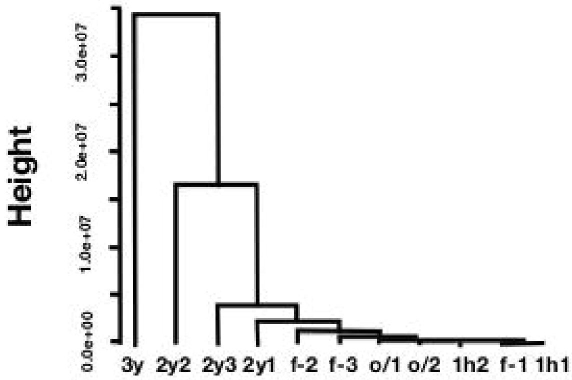
Group Average



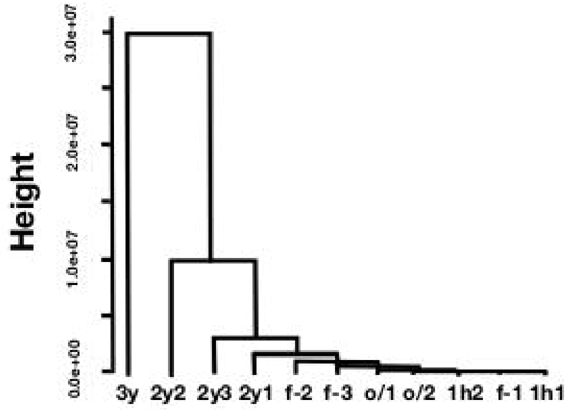
Weighted Average



Centroid



Median



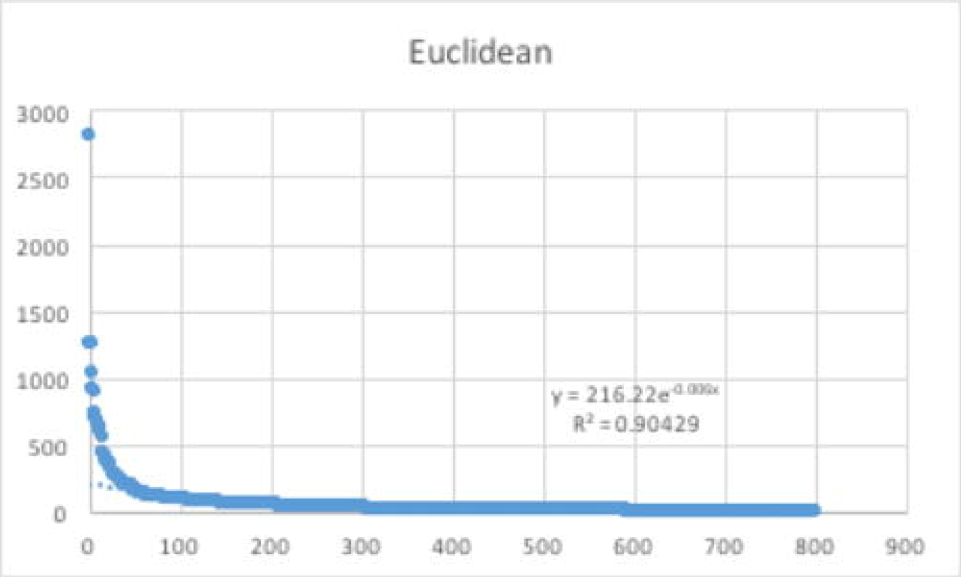
Ward's method

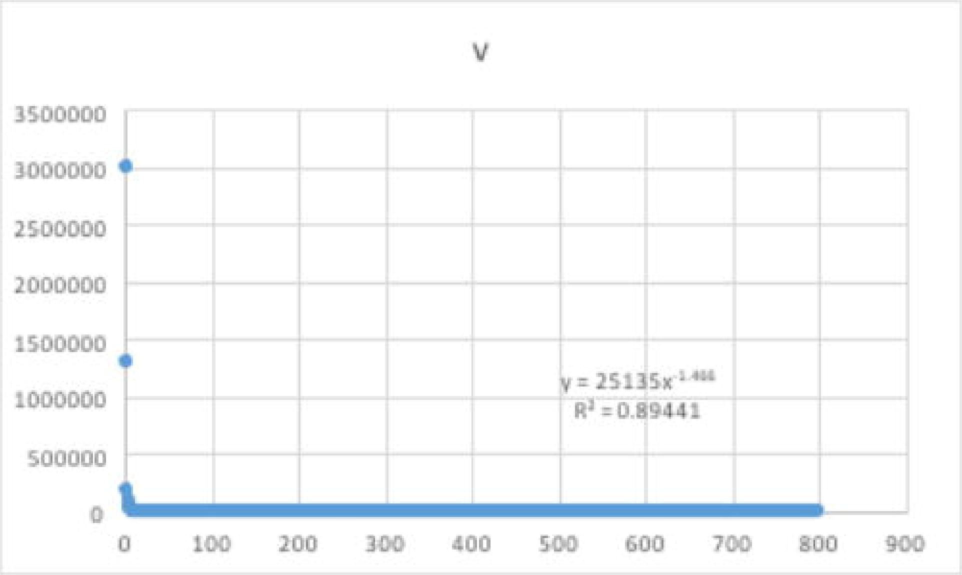
k = 5 means

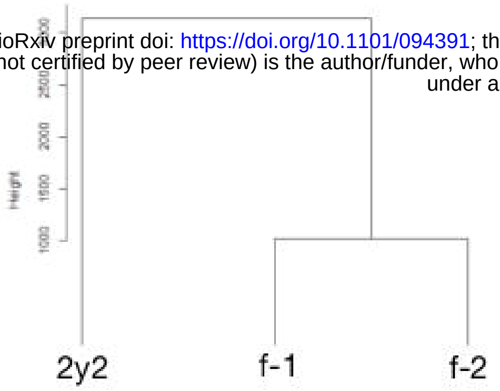
f-1	f-2	f-3	1h1	1h2	o/1	o/2	2y1	2y2	2y3	3y
1	1	1	1	1	1	1	3	2	5	4

Neural Network

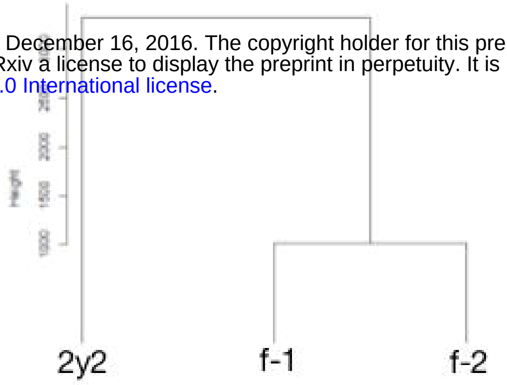
f-1	f-2	f-3	1h1	1h2	o/1	o/2	2y1	2y2	2y3	3y
8.633	9.604	9.360	10.15	8.391	9.151	9.279	-0.220	5.011	6.004	7.896


Euclidean
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
900
$y = 216.22e^{-0.006x}$
$R^2 = 0.90429$

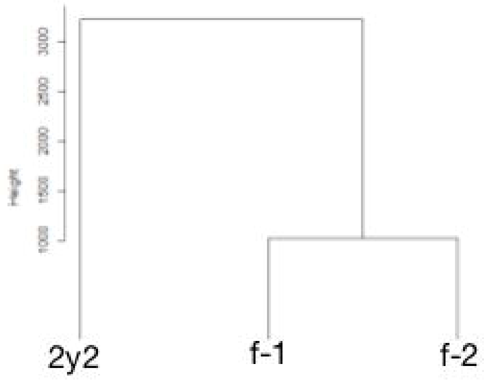

V
3500000
3000000
2500000
2000000
1500000
1000000
500000
0
0
100
200
300
400
500
600
700
800
900
$y = 25135x^{-1.466}$
$R^2 = 0.89441$



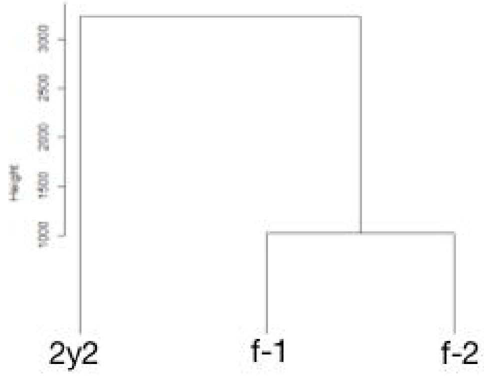
Single Linkage
hclust (*, "single")



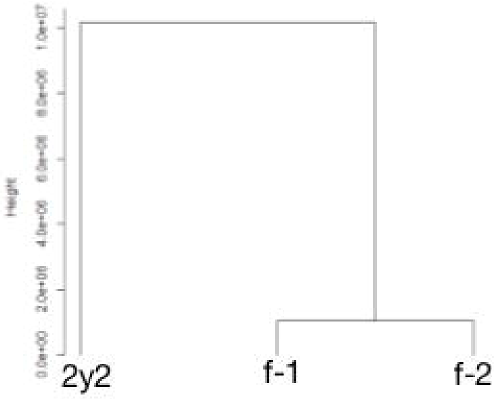
Complete Linkage
hclust (*, "complete")



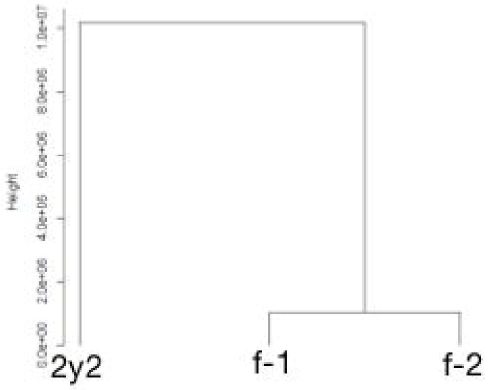
Group Average
hclust (*, "average")



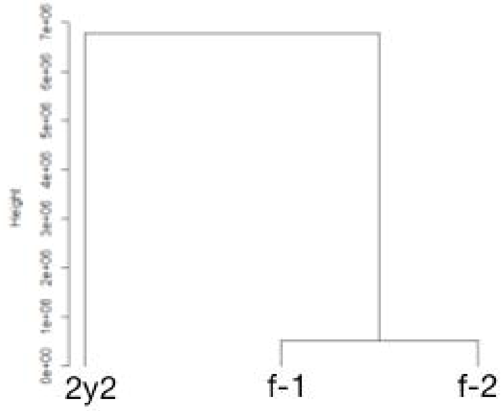
Weighted Average
hclust (*, "mcquitty")



Centroid
hclust (*, "centroid")



Median
hclust (*, "median")



Ward's method
hclust (*, "ward.D")

k = 2 means

f-1	f-2	2y2
1	1	2

Neural Network

f-1	f-2	2y2
8.633	9.604	5.011