Memetic Computing manuscript No.
(will be inserted by the editor)

Manifold Construction Based on Local Distance Invariance

Wei-Chen Cheng · Cheng-Yuan Liou*

Received: date / Accepted: date

Abstract This paper presents a distance invariant manifold that preserves the neighborhood relations among data patterns. All patterns have their corresponding cells in the manifold space. The constellation of neighborhood cells closely resembles that of patterns. The manifold is invariant under the translation, rotation and scale of the pattern coordinates. The neighborhood relations among cells are adjusted and improved in each iteration according to the reduction of the distance preservation energy.

Keywords information visualization · self-organizing map · manifold construction · horizontal gene transfer · economic state · phylogenetic tree · influenza A virus

1 Introduction

Dimension reduction (Kohonen, 1982)(Liou and Musicus, 1990) in the manifold space can display meaningful relationships among patterns. Foundations for various data manifolds have been set down for factorial components (Liou and Musicus, 1990) and for generalized adaline (Wu et al., 2006). They have been successfully applied in various temporal data analyses (Wu et al., 2005).

The principal component analysis (PCA) and multidimensional scaling (MDS) (Torgerson, 1952) are well established linear models that have been developed for such reduction. Nonlinear reduction algorithms have been devised with varying degrees of success. The Isomap (Tenenbaum et al., 2000) and the conformal C-Isomap (de Silva and Tenenbaum, 2002) extend MDS by using the geodesic distance to construct the nonlinear manifold. The Locally Linear Embedding (LLE) (Roweis and Saul, 2000) computes certain linear model coefficients to maintain the local geometric properties

Cheng-Yuan Liou
Department of Computer Science and Information Engineering, National Taiwan University, Republic of China
Tel.: +886233664888 ext. 515
E-mail: cyliou@csie.ntu.edu.tw

Wei-Chen Cheng
Department of Computer Science and Information Engineering, National Taiwan University, Republic of China

in the manifold. Both Isomap and LLE have distinguished performances. Isomap has been extended to find the intrinsic curvature manifold (de Silva and Tenenbaum, 2002), such as a fishbowl surface.

Distortion analysis (Luttrell, 1991; Kohonen, 1999; Liou and Tai, 2000) has been introduced to study the formation mechanism of the self organizing map (SOM) (Kohonen, 1982). The lack of a precise energy function for the SOM is studied in (Erwin et al., 1992). The ill-posed problem of SOM is conjectured by Kohonen, refer the Preface in (Kohonen, 2001). It is very difficult to use the SOM for the non-vector data (Kohonen and Somervuo, 1998). This paper devised a distance preservation energy to construct the manifold. It is a precise energy for the coordinate transformation of the patterns. The ill-posed problem does not exist for such a precise energy function. As this energy is used to manipulate the pattern coordinates, there is no probability manipulation in the manifold construction. The manifold is developed for data visualization only. It is not designed for LVQ or for the clustering purpose.

This energy is devised to display the local relations among patterns directly. Since it uses the relative distances among local patterns to construct the manifold, it is invariant under the translation, rotation and scale of the pattern coordinate. Many existing manifolds are heavily sensitive to the setting of coordinates and obtain serious unreliable results, for example, the eigenvector system and LLE. This kind of invariance is very useful in many applications. The manually assigned numerical codes for the physical entities of patterns are usually arbitrary and abstract. The absolute values of these codes are meaningless. We expect that the distance between two neighborhood patterns carries rich and reliable meaning. The distance preservation manifold is capable of mapping the whole distribution of very high dimensional patterns to a perceptible space, 2D or 3D. We show experiments on market states (Deboeck and Kohonen, 1998; Liou and Kuo, 2002) and influenza protein sequences.

2 Method

Suppose there are P patterns distributed in a D-dimensional pattern space, $X = \{\mathbf{x}^j, \ j = 1, ..., P\}$. Each pattern, $\mathbf{x}^j$, is a D-dimensional column vector and has a corresponding mapped cell, $\mathbf{y}^j$, in the manifold space. The positions of the cells in the manifold space are $Y = \{\mathbf{y}^j, \ j = 1, ..., P\}$. Each cell position, $\mathbf{y}^j$, is a M-dimensional column vector. In the pattern space, giving a pattern $\mathbf{x}^p$, the set of those patterns whose distances to $\mathbf{x}^p$ are less than r are included in the set $U(p, r)$, where r denotes the radius of neighborhood region in the pattern space. The notation $|U(p, r)|$ denotes the number of patterns in the set $U(p, r)$.

Note that the space, M, is a pre-designed space that is continuous without fixed borders. One may set a different M in a different application. This preset M can drastically simplify the manifold problem. All other manifold methods, except SOM, attempt to seek such a space, for example, a wrapped surface in the pattern space.

Consider the local distance invariant manifold (LDIM) energy (Liou et al., 2000),

$$E(r) = \sum_p E^p(r) = \frac{1}{4} \sum_p \sum_{\mathbf{x}^q \in U(p,r)} \left(\left\| \mathbf{y}^p - \mathbf{y}^q \right\|^2 - \left\| \mathbf{x}^p - \mathbf{x}^q \right\|^2 \right)^2 . \qquad (1)$$

The energy is a function of $\mathbf{y}$ and r. The main difference between this energy (1) and that of MDS (Torgerson, 1952) is that (1) is a function of r. All patterns are used in

(1) when r is very large. Few neighbors are used when r is small. These near neighbors are much reliable for setting and supporting, collectively, the position of the cell $\mathbf{y}^p$ in the manifold space. The algorithm for the LDIM that adjusts the cell position, $\mathbf{y}^j$, in the manifold space is as follows.

LDIM Relaxation Algorithm

```
1. Initialize the cell set $Y$.
2. Assign a value to $r$.
3. For each epoch $t$ from $t_0$ to $t_1$
4.     For every pair patterns $\mathbf{x}^p$ and $\mathbf{x}^q$, adjust their cell positions by
```

$$\begin{aligned} \mathbf{y}^p(t) &= \mathbf{y}^p(t-1) - \eta \left[\frac{\partial E(r)}{\partial \mathbf{y}^p} \right]_{t-1}, \\ \mathbf{y}^q(t) &= \mathbf{y}^q(t-1) - \eta \left[\frac{\partial E(r)}{\partial \mathbf{y}^q} \right]_{t-1}. \end{aligned} \tag{2}$$

```
5.     Reduce $r$.
6. End For
```

In the above algorithm, η is the learning rate. The computational complexity for calculating the pairwise distance is $O\left(DP^2\right)$ and finding the neighbors for every pattern is $O\left(P^2\right)$. The computational complexity of Isomap is $O\left(P^3\right)$. It is higher than that of LLE. M is a small number, $M = 2$ or 3 in many cases, in the dimension reduction.

To explain the idea of this algorithm, one may imagine that there are P labeled balls (cells) on a flat table. The surface of this table resembles the low dimensional space, $M = 2$. These balls are free to move on the table. They are confined on this M-dimensional space. Each ball is labeled with its corresponding pattern $\mathbf{x}^p$. The position of the pth ball on the table is $\mathbf{y}^p$. The algorithm seeks a ball distribution Y in M that can resemble the neighborhood relations of X in D. The system energy (1) exerts an implicit, distant and remote, force to the current distribution Y to force it maximally similar (resemble) to the distribution X. One may construct a Gibbs type system for the energy (1) to relax the table distribution as a whole and solve the distribution Y. Here, we prefer a batch operation, Step 4, by using the forces to redistribute the balls.

There are many ways to assign the initial positions of cells in Step 1. For example, one may apply the linear projection methods, such as PCA and MDS, to project all patterns onto the M-dimensional space. Other developed methods, LLE and Isomap, can also be used for the initial assignment. Using the manifold by LLE or Isomap as the initial arrangement is computationally expensive for certain applications.

In Step 2, r denotes the neighborhood radius of the hypersphere in the pattern space. The patterns inside the hypersphere are used in the energy (1). During the end of the relaxation process, the value of r is shrunk to the minimum distance among all pattern pairs. We set the initial value r to be the maximum distance among all pattern pairs and reduce r linearly or exponentially. The computational complexity of updates is

$$\text{(number of iterations)} \times O\left(MP \max_{\mathbf{x}^p} |U(p,r)| \right). \tag{3}$$

This algorithm is different from SOM. The LDIM organizes the positions of the cells based on the distance relations instead of the absolute pattern vectors used in the self-organizing map (SOM) (Kohonen, 1988). These relative distances are readily available

Table 1 Differences between SOM and LDI.

	SOM	LDIM
Number of neurons N	$P >> N$	$P = N$
Lyapunov function	N.A.	E (1)
Neighborhood (N_c)	Isotropic in manifold space	Anisotropic in M
Constellations of neurons	Neurons are fixed in grid positions on a confined rigid plate.	Cells move freely in a continuous space M. M is not a confined space with fixed border.
Clustering	Clustering with various competition laws	N.A.
Coordinate	Sensitive and difficult for non-vector data	Insensitive and invariant

Table 2 Comparison of Energy Function

Method	**Energy function**
LDIM	$E(r) = \frac{1}{4}\sum\limits_{p}\sum\limits_{\mathbf{x}^q \in U(p,r)} \left(\lVert\mathbf{y}^p - \mathbf{y}^q\rVert^2 - \lVert\mathbf{x}^p - \mathbf{x}^q\rVert^2\right)^2$
Isomap	$E = \sum\limits_{p}\sum\limits_{q} \left\lVert \lVert\mathbf{y}^p - \mathbf{y}^q\rVert^2 - \left[d_G\left(\mathbf{x}^p, \mathbf{x}^q\right)\right]^2 \right\rVert$
LLE	$E(W) = \sum\limits_{i} \left\lVert \mathbf{x}^i - \sum\limits_{j} W_{ij}\mathbf{x}^j \right\rVert$

in many applications. The positions of the cells will not be fixed in regular positions as those in SOM. There is no synaptic weight attached to each cell that indicates its constellation in the pattern space. The differences between SOM and LDIM are listed in Table 1. Table 2 lists the energy functions of LDIM, Isomap and LLE. In Table 2, the term, $d_G\left(\mathbf{x}^p, \mathbf{x}^q\right)$, computes the shortest path distance between $\mathbf{x}^p$ and $\mathbf{x}^q$ in a weighted graph G. The weight, W_{ij}, summarizes the contribution of the point $\mathbf{x}^j$ to the $\mathbf{x}^i$. The pseudocode of LDIM is contained in Table 3. In the pseudocode, the parameter r is reduced exponentially. Alternatively, r can be decreased linearly,

$$r(t) \longleftarrow \hat{r} - (\hat{r} - \check{r}) \times \frac{(t - t_0)}{(t_1 - t_0)}. \tag{4}$$

All distance relations will be included in the tuning of each cell's position in the beginning of the relaxation algorithm. The number of neighbors of a pattern, $|U(i,r)|$, will be reduced to zero, δ^i will approach to zero. This will stop the movement of the cell position on the manifold space and reach the convergence.

3 Experiments on artificial data

3.1 Swiss roll dataset

In this example, we show that the sampling density affects the LLE manifold. The swiss roll equation (Liou and Cheng, 2008) is

$$\left(\sqrt{\frac{u}{2\pi}}\sin(3\pi u), \sqrt{\frac{u}{2\pi}}\cos(3\pi u), v\right); \; 0 \le u \le 1 \text{ and } -\frac{3}{10} \le v \le \frac{3}{10}. \tag{5}$$

We uniformly sample data points as patterns along the variable v in the range $\left[-\frac{3}{10}, \frac{3}{10}\right]$ and non-uniformly along u. Let $\mathbf{r}(u)$ be the equation of the curve,

$$\mathbf{r}(u) = \left(\sqrt{\frac{u}{2\pi}}\sin(3\pi u), \sqrt{\frac{u}{2\pi}}\cos(3\pi u)\right); \; 0 \le u \le 1. \tag{6}$$

Table 3 Pseudocode of LDI algorithm

```
For $i = 1$ to $P$
  For $j = 1$ to $M$
    initially assign $y_j^i$ by MDS
  End For
End For

Create distance matrix $S$ by $s_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$

$\check{r} \longleftarrow \min_{i<j} s_{ij}$
$\hat{r} \longleftarrow \max_{i<j} s_{ij}$

For epochs $t = t_0$ to $t_1$
  $r(t) \longleftarrow \check{r} + (\hat{r} - \check{r}) \times e^{-\frac{4(t-t_0)}{(t_1-t_0)}}$
  For $i = 1$ to $P$
      $\delta^i = \sum_{\mathbf{x}^j \in U(i,r(t))} \left( \|\mathbf{y}^i - \mathbf{y}^j\|^2 - (s_{ij})^2 \right) (\mathbf{y}^i - \mathbf{y}^j)$
      $\mathbf{y}^i \longleftarrow \mathbf{y}^i - \eta\delta^i$
  End For
End For
```

The arc length of $\mathbf{r}(u)$ is

$$f(u) = \int_0^u \sqrt{\frac{\partial \mathbf{r}(u')}{\partial u'} \cdot \frac{\partial \mathbf{r}(u')}{\partial u'}} du', \tag{7}$$

where $f(u)$ is the arc length with respect to u. We sample along $f(u)$ and use f^{-1} to calculate the u value. With this u value and the v, we can find a corresponding point on the plane by using the roll formula (5). Using this sampling technique, the probability densities of points may be uniform or non-uniform along the arc length. We can design any sampling densities along the variables, u and v.

Figure 1 shows the manifolds obtained by different algorithms using different density distributions. The left diagrams plot the density distributions with respect to $f(u)$. In these diagrams, the horizontal coordinate, x-axis, is $f(u)$, and the vertical coordinate, y-axis, is the density. Figure 1(a) shows the manifolds of an unbalanced sampling. The blue area of the roll has dense sampling points in unit area and the red area has low density of points. Figure 1(b) shows the manifolds where the red part of the roll has high density. Figure 1(c) shows the manifolds where the density is even along the arc length. Figure 1(d) shows the manifolds where the yellow area (middle portion) has high density.

From Figure 1, we see that the Isomap is not affected much by the density distributions. The sinusoid curves in the LLE manifolds will fluctuate with the densities. The LDIM is not affected by the density distributions. This is because every pattern has its correspondent cell in the manifold space. The number of cells is equals to the number of patterns. The probability density of each pattern is preserved and equal to the density of its cell. The LDIM algorithm is a coordinate transformation algorithm. It maps all events (patterns) from the event space (the pattern space) to their corresponding cells in the manifold space. This algorithm accomplishes the coordinate transformation from the pattern space to the manifold space automatically. The LDIM is a precise energy for such transformation. There is no manipulation on the probability density function in the construction of the manifold. The ill-posed problem in the Preface in Kohonen's book (Kohonen, 2001) does not exist in the LDIM. The LDIM has a perfect energy function and bypasses the problem skillfully. Note that the conformal self-organizing map (Liou and Tai, 1999) is also devised without this problem.

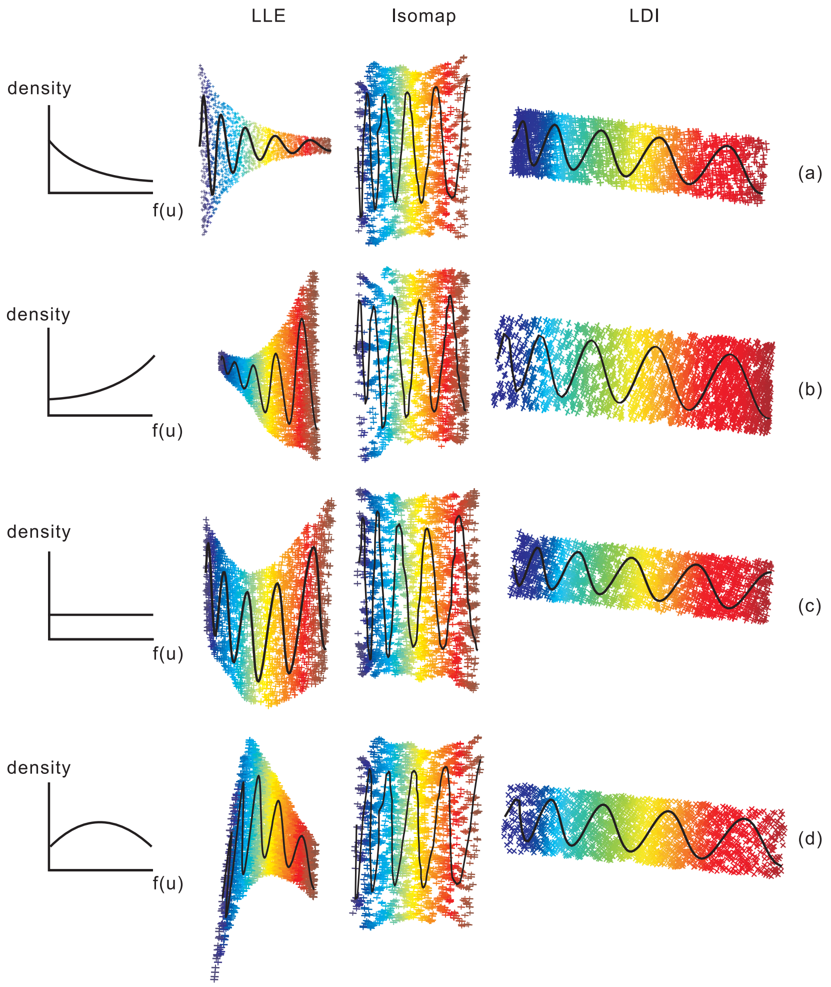

Fig. 1 The manifolds of LLE, Isomap and LDI for different sampling densities along arc length $f(u)$ plotted in the left column.

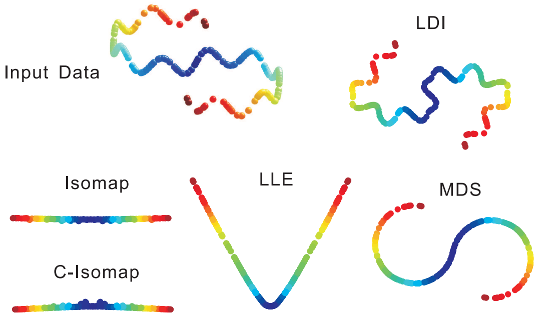

Fig. 2 The top-left image is the pattern curve which has two kinds of structures. The curve has an S-shape on the xy-plane and a sinusoid curve along the z-axis. The manifolds obtained by using LDI, Isomap, C-Isomap, LLE and MDS are plotted with their labels.

3.2 S-curve

The LDIM fully utilizes the manifold space to maintain both global and local structures. As an example, the curve pattern in Figure 2 has two kinds of structures. It consists of 240 pattern points. These points globally form a S-curve on the xy-plane and locally form a sinusoidal curve along the z-axis. The points that have the same color are from the same pattern data. The LDIM reveals these two kinds of structures. Both Isomap with the parameter, $K = 7$, and LLE with the parameter, $K = 12$, derive incorrect structures. MDS maps the points onto a projection plane and reveals the S-curve structure only.

This curve consists of 240 points and there are $(240 \times 239)/2 = 28680$ distances. We divide all 28680 distances into frequency bins (groups). The length of each bin interval is 0.1. We plot the tabulated frequency in each interval, see Figure 3. All distances are in the range $[0, 4.1)$. The notation $[0, 4.1)$ means that the interval includes 0 but exclude 4.1. The group which has the highest frequency is in the interval $[2.0, 2.1)$. There are 2356 distances belong to this interval. The sinusoidal wave induces a small peak in the region $[0.4, 1.1)$. The MDS manifold does not derive the sinusoidal wave. From Figure 3, the MDS histogram is lack of a peak in the region $[0.4, 1.1)$. MDS obtains the detailed frequency information for large distances, $[2, 4.1)$. The Isomap and C-Isomap obtain distorted histogram informations. The LLE histogram is distorted.

4 Experiments on real data

4.1 Economic data

Data reduction techniques are extensively applied in analyzing the economic states (Deboeck and Kohonen, 1998; Liou and Kuo, 2002)(Aranha and Iba, 2009). We use the LDIM to display the states of the global economy. We select $D = 18$ country indices collected from January, 2000 to September, 2009. The index names and country names are listed in Table 4. These 18 indices can reveal certain national policies. We use the monthly data in the analysis. Each pattern vector, $\mathbf{x}^p$, contains the 18 records of the normalized indices of the pth month and is regarded as the month state of the economy.

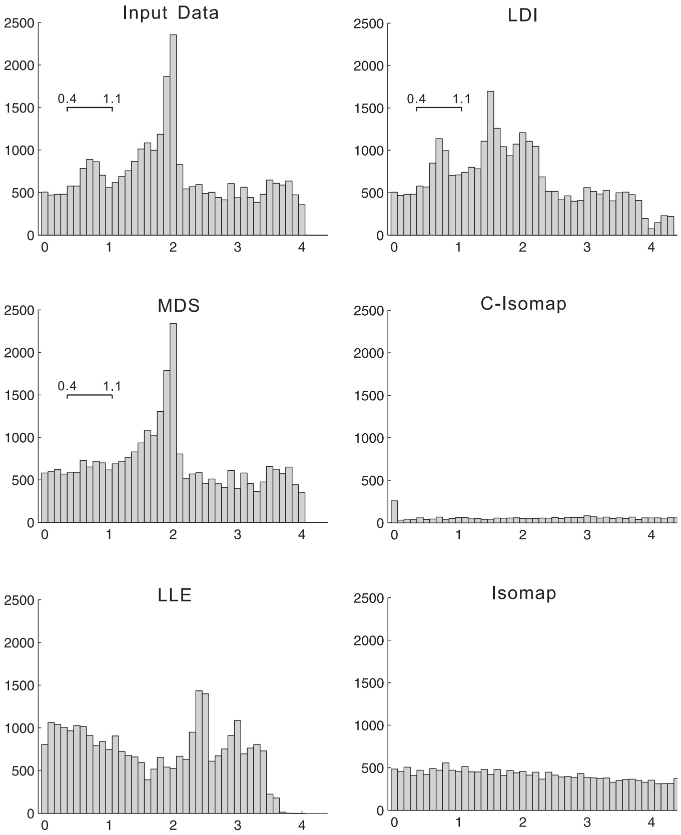

Fig. 3 The histogram of the curve pattern and the manifolds by the five methods.

Each country indices I^p is normalized by

$$\frac{I^p - I^{(January,2000)}}{\left\| I^{(January,2000)} \right\|}. \tag{8}$$

The normalization uses the country indices in January, 2000 as the base. There are $P = 117$ states and each state is an 18-dimension vector ($D = 18$). Figure 4 displays the LDIM ($M = 2$). The black solid circle denotes the state cell in January 2000 and the concentric circles denote the cell in September, 2009. The green line shows the LDIM when r is set to the maximum distance, $\hat{r} = 6.474$. The red line is the converged manifold when r is decreased to the minimum distance, $\check{r} = 0.070$. We record the relaxation processes between the green line and the red line, from light gray to dark gray. In Figure 4, the red line reveals much more detailed information than that of the green line. It clearly shows that the subprime mortgage mess in July, 2007 is right on the turning point of the global economy.

Figure 5 displays the 3D manifold ($M = 3$). The line is the converged manifold. The computation time is 1524 seconds. With this figure, one see the significant market trends over time vividly. Both Figure 4 and Figure 5 show that the subprime mortgage mess in July 2007 is the turning point of the global economy. The 3D manifold provides similar trend information as that of 2D manifold.

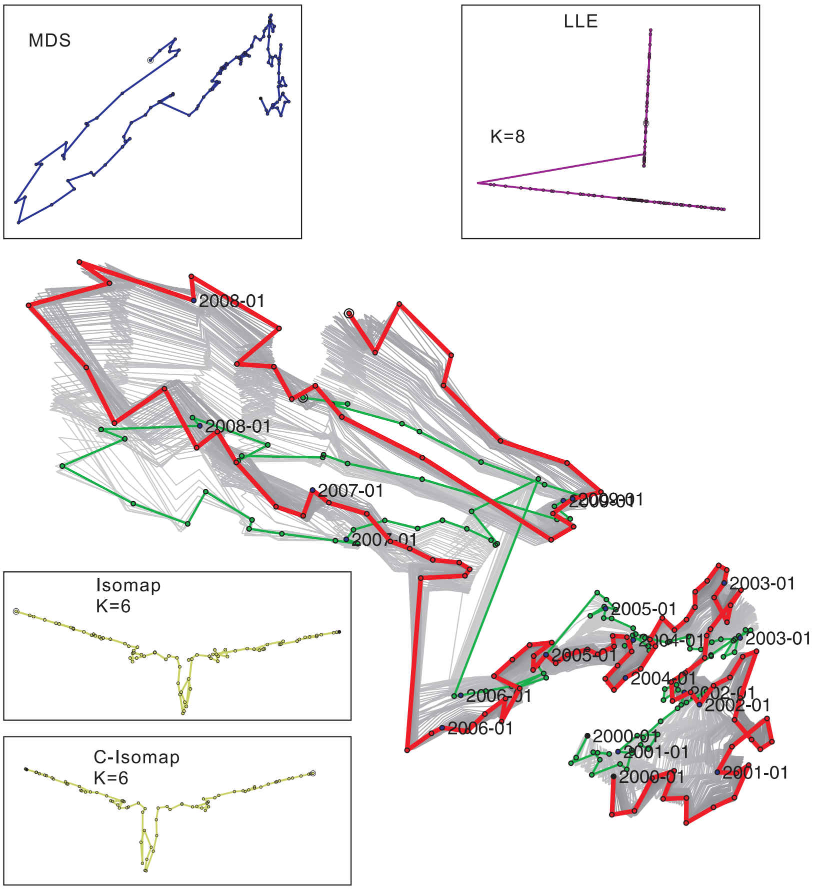

Fig. 4 The green line plots the initial setting of the LDI manifold. Gray lines, from light to dark, record the results during relaxation epochs. The red line plots the converged LDI manifold. We also plot the manifolds by Isomap, C-Isomap, LLE and MDS.

Table 4 World market indices

Amsterdam	Australia	Bombay	Frankfurt	New York
AEX	ALL ORDS	BSE SENSEX	DAX	DJ-INDUS
London	Hong Kong	Jakarta	Kuala	Korea
FTSE100	HANG SENG	JKSE	KLSE	KOSPI
Nasdaq	Osaka	New York	OSLO	New York
NASDAQ	NIKKEI 225	NYSE COMP	OBX	S&P 500
Shanghai	Swiss	Taiwan		
SSE	SWISS MARKET	TAIEX		

We list several important events in Table 5 that have great impact to the global economic system and may have long-term effects. The manifold shows that the economic states before the year 2006 are different from the states after the year 2007. They are two very different patterns. Those states after the year 2007 are very difficult to predict from those before the year 2006. There is a sharp turn on the October, 2007 that is three months after the reveal of the mess news. This turn shows the trend deviation

Table 5 Events that have great impact to the economics.

Time	**Events**
1978–1982	Second oil crisis
Oct., 1987	New York stock crash
Aug., 1990	The Gulf war
Dec., 1997	Asian Economic Storm
Sept., 2001	US 911 (Economic deterioration)
Nov., 2002–Jul., 2003	Outbreak of SARS
Nov., 2003–Feb., 2004	Avian influenza
Dec., 2004	Indian ocean tsunami
Aug., 2005	Hurricane Katrina
2006	Drought in Australia
2007	Subprime mortgage crisis
Jan., 2008	Chinese winter storms
May., 2008	Sichuan earthquake
April, 2009	H1N1 swine flu

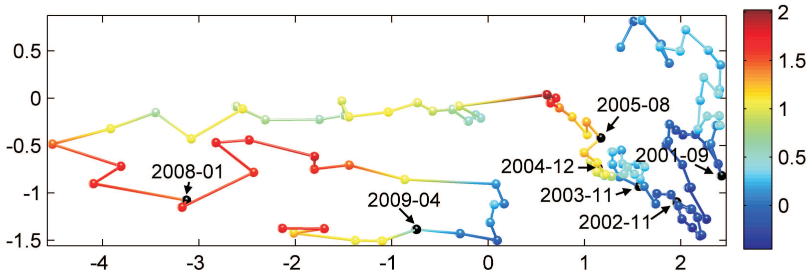

Fig. 5 The global markets which evolve over time. The manifold are on the three dimensional space and the third dimension is represented by the colors.

of the states. Extremely long distances and sharp turns between two successive states indicate the rough situations, such as the large drift during September 2008. This figure is suitable for many visualization purposes and facilitates various interpretations on the content of the world states.

One can apply the manifold technique to interpolate the incomplete data records and extrapolate the predicted states. As for the interpolation of missing data, suppose there are three missing records during certain month. We use all $P-1$ available data to construct the manifold with $P-1$ cells. These $P-1$ cell positions are then fixed. Then, we insert one new cell in the manifold for the incomplete month and train its position using the rest $15 = 18 - 3$ records in the vectors, $\mathbf{x}^p$, of all P states. Note that the rest $P(P-1)/2$ distances are calculated by using the 15 rest records of all P vectors. This inserted cell will converge to a new position and we obtain a fixed manifold with P cells. The distances of those near neighbors of this inserted cell in the manifold can be used as the weights to interpolate the missing three records. By a weighted fitting of these three missing records among the neighbors, one can estimate the missing three records. With the state trend and distances among neighbors, the extrapolation of the predicted state can be accomplished in a similar way.

We also select five indices, GDP growth, consumer price index, import value index, export value index, and unemployment rate to display the individual national economic states. The data source is recorded in the publications by WDI (World Development Indicators) of World Bank. There are 21 years from the year 1987 to the year 2007. Each

Table 6 The settings of parameter r.

	$\check{r}$	$\hat{r}$
US	0.126	3.932
GB	0.260	3.622
JP	0.281	3.405
CN	0.079	3.896

data, $\mathbf{x}^p$, is a five dimensional column vector and contains the five indices. Each index is normalized to the range $[-1, 1]$, separately. The 210 = $(21 \times 20)/2$ pair distances among the twenty-one year vectors are calculated and used in the LDIM algorithm. We set the learning rate $\eta = 0.01$. The algorithm is operated with 100 millions epochs. The parameter r is exponentially decreased. The maximum $\hat{r}$ and minimum $\check{r}$ are listed in Table 6. Figure 6 shows the 3D manifolds for four countries. They have different market patterns and trends. We mark several abnormal states. The results in (Liou and Kuo, 2002) are consistent with those in Figure 6.

4.2 Manifolds for H5N1 and H1N1 proteins

Due to the horizontal gene transfer (Meuth et al., 2009), the life tree may not be an adequate structure to fully display the evolution relations among generations. Alternatively, we show how to display the generations in the manifold space.

The H1N1 and H5N1 are subtypes of the Influenza A virus which can cause illness in humans and many animal species. The H5N1 subtype has been reported in 445 human cases and has caused 263 human deaths (WHO, 2009). The pathway of H5N1 spread is still unclear.

We use the protein HA segments of all available DNA sequences saved in NCBI's Influenza Virus Resource (Bao et al., 2008) . We select 184 distinct Influenza A(H1N1) protein sequences and 196 distinct H5N1 sequences. All 380 sequences are full lengths. The minimum and maximum lengths of the H1N1 sequences are 554 and 567 amino acids respectively. The minimum and maximum lengths of the H5N1 sequences are 552 and 575 amino acids. The host of all the selected isolates is human. All the selected Influenza A(H1N1) sequences are recorded during the year 2009. The H5N1 sequences are recorded during the 13 years from 1997 to 2009.

We align all H1N1 sequences together. Multiple-sequence alignments were performed using the Clustal W2 program (Larkin et al., 2007). We compute the Hamming distances between every two sequences. Fig. 7 shows the 2D manifold for the $P = 184$ H1N1 sequences. Each aligned sequence consists of 567 amino acids. There are $184 \times 183 \div 2 = 16836$ Hamming distances among the 184 sequences. The neighborhood region, r, is reduced from $\hat{r} = 20$ to $\check{r} = 1$ in the algorithm. The algorithm is operated with 500000 epoches. There exists only one cluster. The average center of the cluster is marked with a black square. The two sequences close to the center, $\mathbf{y}_{mean} = \frac{1}{184} \sum_{p=1}^{184} \mathbf{y}^p$, are marked with two black circles. They are the isolate ACQ99610 and the isolate ACR81633. There is a overlap between these two black circles. We may expect that certain cells near the center may be the grandmother of the H1N1 virus. Finding the grandmother cell is useful for many medicine goals, such as the design of vaccine and the trace of the virus source. The evolution trend of the DNA mutations is displayed by colors. One can monitor the sampled isolates in the trend. The three

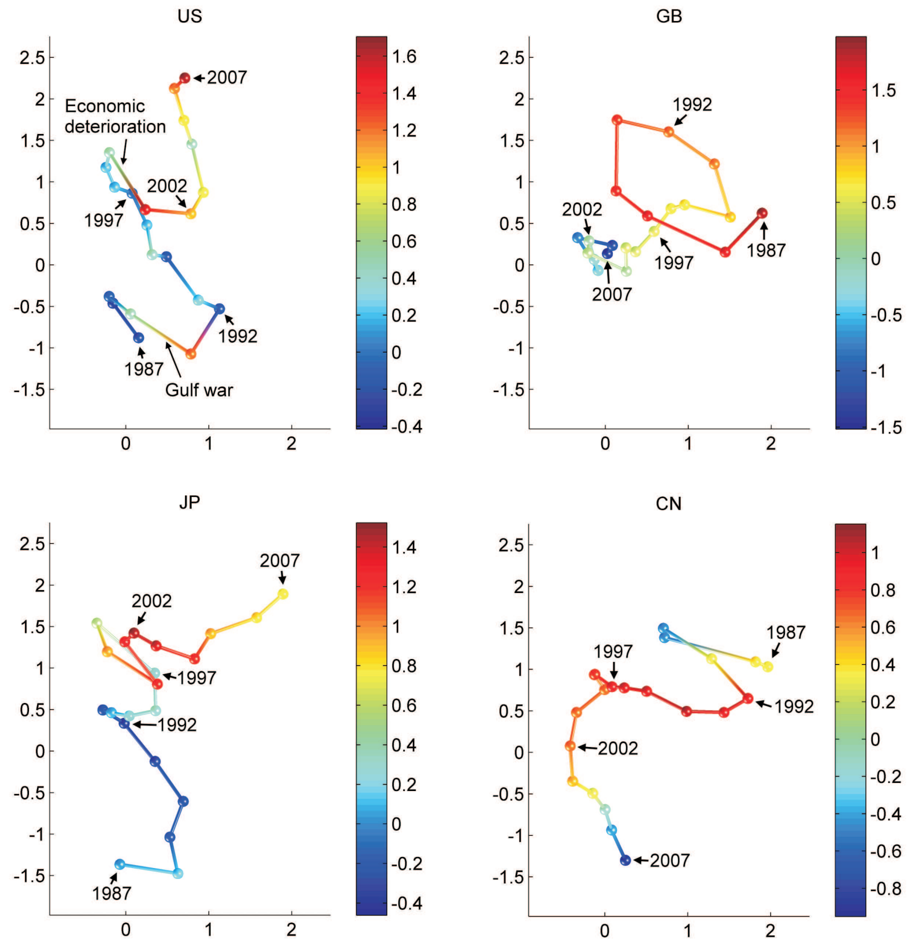

Fig. 6 Economic states of four countries displayed in 3D LDI manifolds. Colors are for the third dimension.

gray ellipses are the one, two and three times of the covariance. Both the covariance and the center are calculated in the 2D manifold space, Y.

Fig. 8 shows the distance invariant manifold for the 196 H5N1 gene sequences. In this case $P = 196$. Each aligned sequence consists of 583 amino acids. There are $(196 \times 195)/2 = 19110$ Hamming distances among these 196 sequences. The neighborhood region, r, is reduced from $\hat{r} = 74$ to $\check{r} = 1$ in the algorithm. The gray ellipses show the covariance information in the manifold space. The sequence close to the center is marked with a black circle. It is the isolate ACA64009.

4.3 Phylogenetic Tree

The manifold space M may have other shapes. We show an application for the manifold space that has a tree like structure. The phylogenetic tree is useful to display the

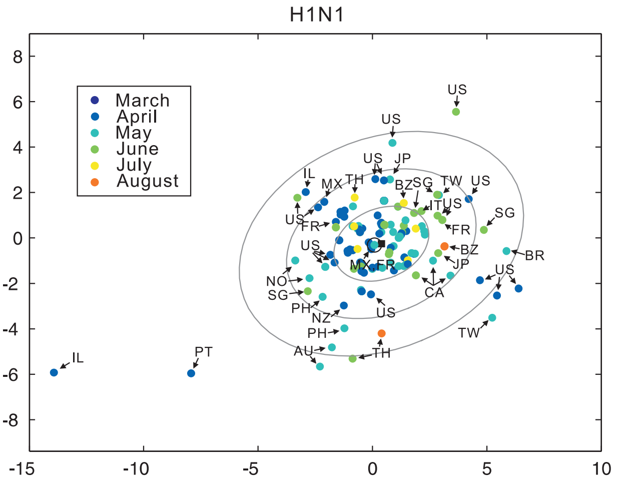

Fig. 7 The 2D LDI manifold for H1N1 gene sequences. The colors mark the monthly information from March to August.

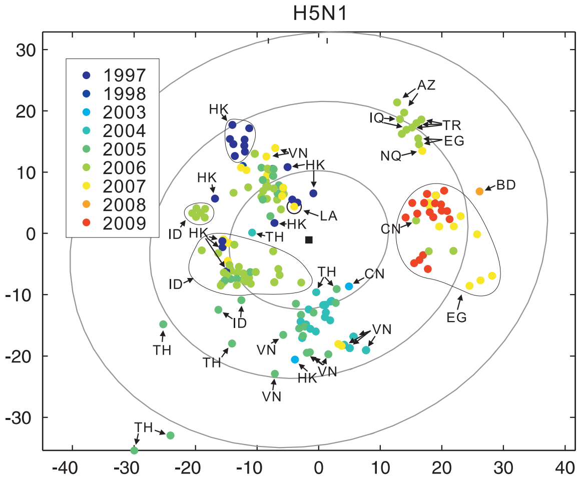

Fig. 8 The 2D LDI manifold for H5N1 gene sequences. The colors mark the yearly information from 1997 to 2009.

inter-relations among species (Johnson, 1967; Sattath and Tversky, 1977; Saitou and Nei, 1987). Usually, the tree is constructed based on the minimization of the overall difference between all path lengths of the tree and their corresponding distances among species that stored in a distance matrix (Farris, 1972). The path distance between two leaf nodes is the sum of the lengths of the branches along the path connecting these two nodes. According to the construction, the sum of all distances of all node pairs is close to that obtained from the distance matrix. Any path length in the tree should be fitted to its corresponding species relation in the matrix. Based on the LDIM, we rewrite the LDIM energy for the tree path estimation to fine-tune its branch lengths for the H5N1 tree and SARS tree.

Given a set of distance relations among species, we plan to construct the branch lengths of a given tree that meet the distance relations with its path lengths. The tree is an undirected binary tree. Its branches have no direction. Suppose there are total P species, $\{\mathbf{x}^p, p = 1, ..., P\}$, where $\mathbf{x}^p$ is a column vector that saves the amino acid sequence of the pth species. The tree has P leaf nodes and $2P-2$ branches. The tree structure can be saved in a matrix, $A_{(P(P-1)/2)\ by\ (2P-2)}$. In A, the row indices denote the tree paths corresponding to every species pairs, $\{\left(\mathbf{x}^i, \mathbf{x}^j\right)$, $i = 1, ..., P$; $j = i+1, ..., P\}$, and the column indices denote the $2P-2$ branches of the tree. The element, a_{ij}, of the matrix A is set as

$$a_{ij} = \begin{cases} 0 \text{ , when the path } i \text{ doesn't contain the branch } j. \\ 1 \text{ , when the path } i \text{ includes the branch } j. \end{cases} . \tag{9}$$

Let $\delta = \left[\left\|\mathbf{x}^1 - \mathbf{x}^2\right\|, \left\|\mathbf{x}^1 - \mathbf{x}^3\right\|, \ldots, \left\|\mathbf{x}^{P-1} - \mathbf{x}^P\right\|\right]^T$ be a $P(P-1)/2$-by-1 column vector that consists of all distances between species pairs. Assume that the lengths of the $2P-2$ branches are variables, $\mathbf{z} = \left[z_1, \ldots, z_{(2P-2)}\right]^T$. Then seek a solution for $\mathbf{z}$,

$$A\mathbf{z} \approx \delta, \text{ subject to } \mathbf{z} \geq \mathbf{0}. \tag{10}$$

The method in (Sattath and Tversky, 1977) suggested using the least square method with non-negative constraint to modify the branch lengths of the tree. We rewrite the LDIM energy to solve the variables $\mathbf{z}$, $\hat{E}(r)$,

$$\hat{E}(r) = \sum_p \sum_{\mathbf{x}^q \in U(p,r)} \left(t(p,q) - \|\mathbf{x}^p - \mathbf{x}^q\|\right)^2 \tag{11}$$

where $t(p,q)$ denotes the path length from the leaf node p to the node q. $U(p,r)$ is the set that contains all neighbors $\mathbf{x}^q$ of $\mathbf{x}^p$ that are within the distance range r; $U = \{\mathbf{x}^q;\ \|\mathbf{x}^p - \mathbf{x}^q\| \leq r\}$. The variables z can be solved by minimizing this energy.

We adjust the branch lengths $\mathbf{z}$ by applying the gradient descent method to the energy $\hat{E}(r)$ and restrict the value of $\mathbf{z}$ to be larger than or equal to zero. The value of r is reduced during the relaxation.

4.4 Tree experiments

Three datasets are used in the experiments. One is Case's data (Case, 1978) that contains the immunological distances among nine frog (Rana) species. The second dataset is the influenza A virus, H5N1 subtype, from NCBI (National Center for Biotechnology Information). The third dataset is the SARS-CoV genome sequences. We will employ the UPGMA method (Unweighted Pair Group Method with Arithmetic mean) by (Sokal and Sneath, 1963) or the neighbor-joining method by (Saitou and Nei, 1987) to build the initial tree, A, and then use the LDIM energy (11) to fine tune the branch lengths.

Figure 9 shows the estimated branch lengths by the LDIM algorithm. The lengths by Sattath (Sattath and Tversky, 1977) are also plotted in this figure for comparison. The LDIM algorithm obtains very different branches for the subtree that contains the five species, R. aurora, R. boylii, R. cascadae, R. muscosa and R. pretiosa. After convergence, we calculate the performance using the formula, $MDI(r)$, for the LDIM

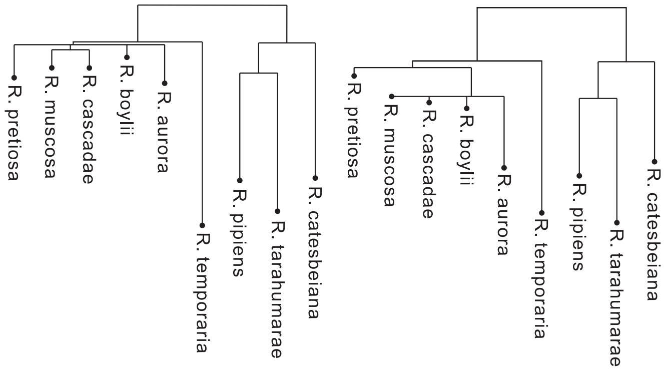

Fig. 9 The trees are constructed by UPGMA. The branch lengths are obtained by LDI algorithm (left) and Sattath's algorithm (right).

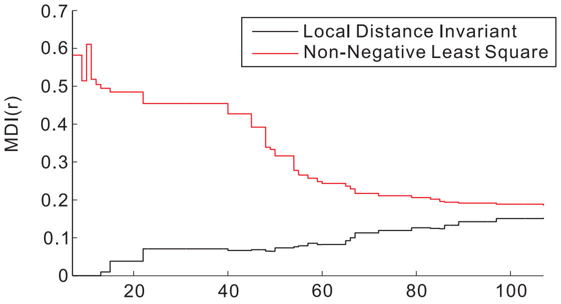

Fig. 10 The performance comparison, MDI (r).

algorithm and Sattath's method. The measurement of distance invariance, $MDI\left(r\right)$, is

$$MDI\left(r\right)=\frac{1}{\sum\limits_{p}\left|U\left(p,r\right)\right|}\sum_{p}\sum_{\mathbf{x}^q\in U\left(p,r\right)}\frac{\sqrt{\left(t\left(p,q\right)-\left\|\mathbf{x}^p-\mathbf{x}^q\right\|\right)^2}}{\left\|\mathbf{x}^p-\mathbf{x}^q\right\|}. \tag{12}$$

The performance is plotted in Figure 10. In this figure, the x-axis denotes r and y-axis denotes the calculated value of $MDI\left(r\right)$ in (12). The performance shows that the LDIM algorithm obtains very precise length information for those small distance species.

The LDIM algorithm is used to construct the phylogenetic tree for the amino acid sequences in the segment one region (PB2) of bird flu, H5N1. The sequences in Influenza Virus Resource (Bao et al., 2008) are used in the construction. All redundant sequences are removed and will not be used. The tree is constructed for the $P = 97$ protein sequences of H5N1 recorded from 1997 to 2007 that hosted only on human. There are $4656 = (97 \times 96)/2$ distances for all sequence pairs. The lengths of the sequences after performing the multiple-alignment, (Edgar, 2004), are all 770. UPGMA uses the Hamming distances among the aligned sequences to build the tree, see Figure 11. The initial branch lengths of the UPGMA tree are obtained by Sattath's method. These lengths are used in the LDIM algorithm as the initial setting. The neighborhood region, r, is reduced from $\hat{r} = 51$ to $\check{r} = 1$ in the algorithm. In Figure 12, the performance (12) shows that the LDIM algorithm obtains very precise lengths for close species.

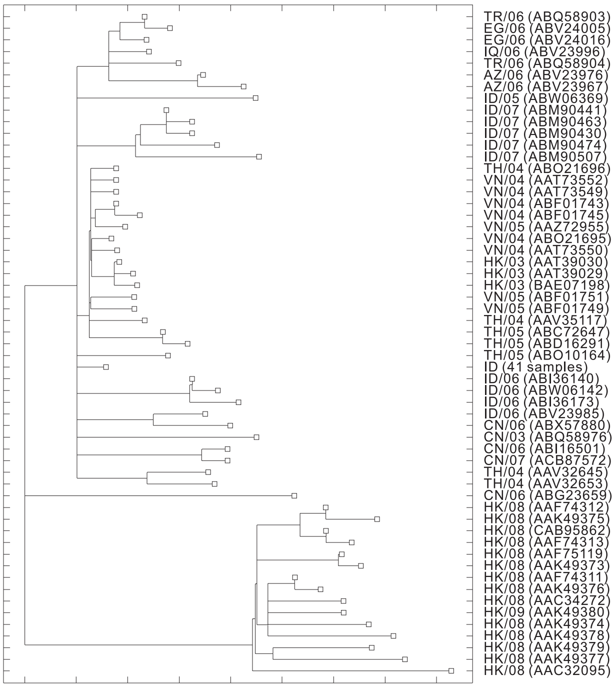

Fig. 11 The initial tree is obtained by UPGMA and the branch lengths are estimated by LDI manifold.

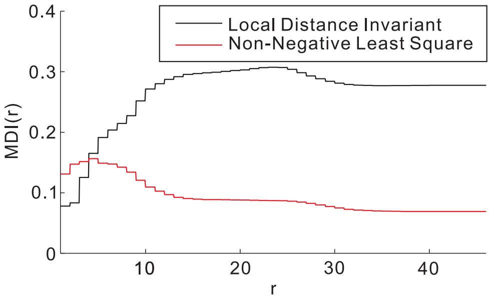

Fig. 12 Performance comparison, MDI (r), on the estimated branch lengths of the H5N1 tree, in Fig. 11, obtained by the LDI algorithm and the non-negative least square algorithm by (Sattath and Tversky, 1977).

We now construct the phylogenetic tree for SARS. (Rota et al., 2003) analyzed the sample, SARS-CoV, under the accession number, AY278741, in Genbank. They applied phylogenetic analysis for the proteins from known coronaviruses and the predicted proteins produced from SARS-CoV. (Marra et al., 2003) studied the SARS genome, named Tor2 (AY274119.3), that consists of 29751 base pairs in Genbank. They showed that the SARS virus does not closely resemble any of the three previously known groups of coronaviruses. Figure 13 shows the results of the LDIM algorithm. The four initial trees are constructed by the neighbor-join method. The results clearly show that SAR-CoV is not in the group of coronavirus.

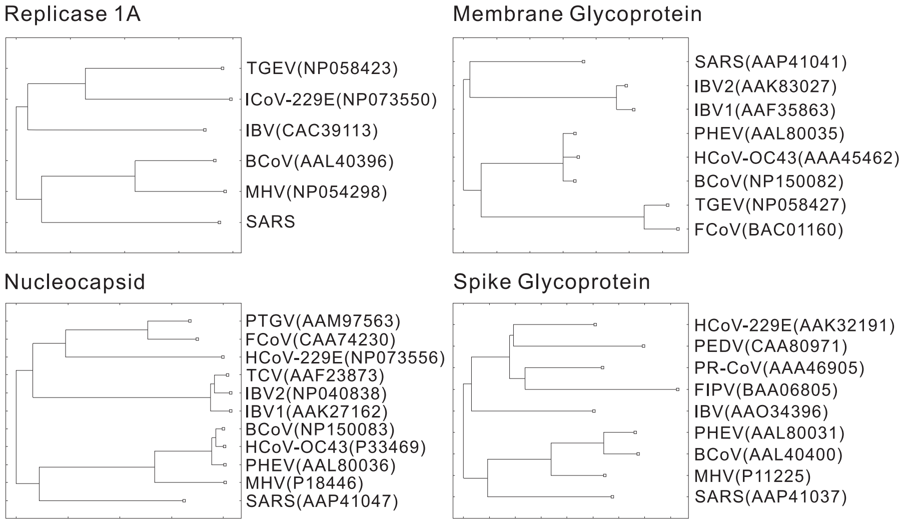

Fig. 13 Branch lengths by LDI algorithm for the (Marra et al., 2003)'s SARS dataset. The four initial trees are obtained by the neighbor-joining method by (Saitou and Nei, 1987).

Finally, we briefly summarize several features of the LDIM. It use the relative distance between patterns to construct the low dimensional manifold. This manifold is invariant under the translation, rotation and scale of pattern coordinates. The constellations of cells are fixed reliably by their neighborhood patterns collectively. The constellation displays both the global and local details of the pattern structure. This manifold is useful in many applications, such as pattern recognitions (Liou and Yang, 1996); time series; chain and tree branch lengths. It is relatively difficult to display chains or tree branches in SOM. The LDIM can be applied to many other invariance preservation problems, such as angular invariance and conformal invariance.

Acknowledgement.

This work was supported by National Science Council NSC 97-2221-E-002-207-MY3.

References

[Aranha and Iba, 2009]Aranha C and Iba H (2009) The Memetic Tree-based Genetic Algorithm and its application to Portfolio Optimization. Memetic Comput 1:139–151

[Bao et al., 2008]Bao Y, Bolotov P, et al. (2008) The Influenza virus resource at the national center for biotechnology information. J Virol 82:596–601

[Case, 1978]Case SM (1978) Biochemical systematics of members of the genus Rana native to western north America. Syst Zool 27:299–311

[de Silva and Tenenbaum, 2002]de Silva V and Tenenbaum JB (2002) Global versus local methods in nonlinear dimensionality reduction. In: Adv Neur Inf Process Syst 15, pp. 705–712

[Deboeck and Kohonen, 1998]Deboeck G and Kohonen T (1998) Visual explorations in finance: with self-organizing maps. Springer

[Edgar, 2004]Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

[Erwin et al., 1992]Erwin E, Obermayer K, and Schulten K (1992) Self-organizing maps: ordering, convergence properties and energy functions. Biol Cybern 67:47–55

[Farris, 1972]Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

[Johnson, 1967]Johnson SC (1967) Hierarchical clustering schemes. Psychometrika 32(3):241–254

[Kohonen, 1982]Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

[Kohonen, 1988]Kohonen T (1988) Self-organization and associative memory. Springer-Verlag, Berlin, ed 2

[Kohonen, 1999]Kohonen T (1999) Comparison of SOM point densities based on different criteria. Neural Comput 11:2081–2095

[Kohonen, 2001]Kohonen T (2001) Self-organizing maps. Springer

[Kohonen and Somervuo, 1998]Kohonen T and Somervuo P (1998) Self-organizing maps of symbol strings. Neurocomputing 21:19–30

[Larkin et al., 2007]Larkin MA, Blackshields G, Brown NP, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

[Liou et al., 2000]Liou C-Y, Chen H-T, and Huang J-C (2000) Separation of internal representations of the hidden layer. In: Proc Int Comput Symp Workshop on Artif Intell, pp. 26–34

[Liou and Cheng, 2008]Liou C-Y and Cheng W-C (2008) Manifold construction by local neighborhood preservation. In: Lect Notes Comput Sci, volume 4985, pp. 683–692

[Liou and Kuo, 2002]Liou C-Y and Kuo Y-T (2002) Economic state indicator on neuronic map. In: Int Conf Neural Inf Process, volume 2, pp. 787–791

[Liou and Musicus, 1990]Liou C-Y and Musicus BR (1990) Separable cross-entropy approach to power spectrum estimation. IEEE T Acoust Speech 38:105–113

[Liou and Tai, 1999]Liou C-Y and Tai W-P (1999) Conformal self-organization for continuity on a feature map. Neural Netw 12:893–905

[Liou and Tai, 2000]Liou C-Y and Tai W-P (2000) Conformality in the self-organization network. Artif Intell 116:265–286

[Liou and Yang, 1996]Liou C-Y and Yang H-C (1996) Handprinted character recognition based on spatial topology distance measurement. IEEE T Pattern Anal 18(9):941–945

[Luttrell, 1991]Luttrell S (1991) Code vector density in topographic mappings: Scalar case. IEEE T Neural Netw 2:427–436

[Marra et al., 2003]Marra MA, Jones SJ, et al. (2003) The Genome sequence of the SARS-associated coronavirus. Sci 300(5624):1399–1404

[Meuth et al., 2009]Meuth R, Lim M-H, Ong Y-S, and Wunsch II DC (2009) A proposition on memes and meta-memes in computing for higher-order learning. Memetic Comput 1:85–100

[Rota et al., 2003]Rota PA, Oberste MS, et al. (2003) Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Sci 300:1394–1399

[Roweis and Saul, 2000]Roweis ST and Saul LK (2000) Nonlinear dimensionality reduction by locally linear Embedding. Sci 290:2323–2326

[Saitou and Nei, 1987]Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

[Sattath and Tversky, 1977]Sattath S and Tversky A (1977) Additive similarity trees Psychometrika 42:319–345

[Sokal and Sneath, 1963]Sokal RR and Sneath PHA (1963) Principles of numerical taxonomy. W. H. Freeman, San Francisco

[Tenenbaum et al., 2000]Tenenbaum J, de Silva V, and Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Sci 290:2319–2323

[Torgerson, 1952]Torgerson WS (1952) Multidimensional scaling, I: Theory and Method. Psychometrika 17:401–419

[WHO, 2009]WHO (2009) Epidemic and pandemic alert and response (EPR). http://www.who.int/csr/disease/avian_influenza/country/cases_table_2009_12_11/en/index.html

[Wu et al., 2006]Wu J-M, Lin Z-H, and Hsu PH (2006) Function approximation using generalized adalines. IEEE T Neural Netw 17:541–558

[Wu et al., 2005]Wu J-M, Lu C-Y, and Liou C-Y (2005) Independent component analysis of correlated neuronal responses in area MT. In: Proc Int Conf Neural Inf Process, pp. 639–642