BIOINFORMATICS

Vol. 26 ECCB 2010, pages i531–i539
doi:10.1093/bioinformatics/btq376

Nonlinear dimension reduction and clustering by Minimum Curvilinearity unfold neuropathic pain and tissue embryological classes

Carlo Vittorio Cannistraci[1,2,3,4,5,*], Timothy Ravasi[1,5], Franco Maria Montevecchi[3], Trey Ideker[5] and Massimo Alessio[2,*]

[1]Red Sea Integrative Systems Biology Lab, Computational Bioscience Research Center, Division of Chemical and Life Sciences and Engineering, King Abdullah University for Science and Technology (KAUST), Jeddah, Kingdom of Saudi Arabia, [2]Proteome Biochemistry, San Raffaele Scientific Institute, via Olgettina 58, 20132 Milan, [3]Department of Mechanics, [4]CMP Group, Microsoft Research, Politecnico di Torino, c/so Duca degli Abruzzi 24, 10129 Turin, Italy, [5]Department of Bioengineering and Department of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093 USA

ABSTRACT

Motivation: Nonlinear small datasets, which are characterized by low numbers of samples and very high numbers of measures, occur frequently in computational biology, and pose problems in their investigation. Unsupervised hybrid-two-phase (H2P) procedures—specifically dimension reduction (DR), coupled with clustering—provide valuable assistance, not only for unsupervised data classification, but also for visualization of the patterns hidden in high-dimensional feature space.

Methods: 'Minimum Curvilinearity' (MC) is a principle that—for small datasets—suggests the approximation of curvilinear sample distances in the feature space by pair-wise distances over their minimum spanning tree (MST), and thus avoids the introduction of any tuning parameter. MC is used to design two novel forms of nonlinear machine learning (NML): Minimum Curvilinear embedding (MCE) for DR, and Minimum Curvilinear affinity propagation (MCAP) for clustering.

Results: Compared with several other unsupervised and supervised algorithms, MCE and MCAP, whether individually or combined in H2P, overcome the limits of classical approaches. High performance was attained in the visualization and classification of: (i) pain patients (proteomic measurements) in peripheral neuropathy; (ii) human organ tissues (genomic transcription factor measurements) on the basis of their embryological origin.

Conclusion: MC provides a valuable framework to estimate nonlinear distances in small datasets. Its extension to large datasets is prefigured for novel NMLs. Classification of neuropathic pain by proteomic profiles offers new insights for future molecular and systems biology characterization of pain. Improvements in tissue embryological classification refine results obtained in an earlier study, and suggest a possible reinterpretation of skin attribution as mesodermal.

Availability: https://sites.google.com/site/carlovittoriocannistraci/home

Contact: kalokagathos.agon@gmail.com; massimo.alessio@hsr.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

1 INTRODUCTION

1.1 The machine learning perspective

Visualization and discrimination as well as supervised and unsupervised classifications are widely employed in computational biology for the investigation and analysis of patterns hidden in wet-lab data. In the literature, 'supervised classification' is frequently simplified into 'classification', and 'unsupervised classification' into 'clustering' and this may give rise to misunderstanding. To avoid terminological ambiguity, 'classification' is adopted throughout this article to describe the general task of sample group attribution, while the issue of whether such attribution is supervised or unsupervised will be specified as and when necessary.

Supervised methods for feature selection and classification present several pitfalls (Smialowski *et al.*, 2009), and small datasets make analysis problematic (Martella, 2006). Complications particularly intensify when samples are nonlinearly related in the high-dimensional feature space obtained from high-throughput genomic and proteomic measures. When the aim is to classify a low number of samples characterized by a very large number of genes, problems with parameter estimation may arise, and dimensional reduction followed by clustering (Martella, 2006) is a valuable response to this scenario. Principal component analysis (PCA) has often been employed (Martella, 2006) in combination with a clustering algorithm that groups homogeneous classes on the basis of principal components, but this approach is insufficiently powerful to deal with nonlinear datasets. In this article, we describe the use of nonlinear hybrid-two-phase (H2P) unsupervised machine learning (ML) methodologies—specifically dimension reduction (DR) in conjunction with clustering—for the concurrent visualization and classification of biological samples. Our aim is to address the issue of nonlinearity and to improve the classification accuracy of recently proposed small nonlinear datasets. The methodological innovation we introduce is a principle called 'Minimum Curvilinearity' (MC), which is used as framework for two novel forms of nonlinear ML (NML): Minimum Curvilinear embedding (MCE) for DR and Minimum Curvilinear affinity propagation (MCAP) for clustering. For small datasets, the 'MC' principle suggests the estimation of curvilinear (geodesic) distances between sample data points as pair-wise distances over their minimum spanning tree (MST) constructed in feature space.

To test efficacy of the proposed algorithms, we considered locally linear embedding (LLE) (Roweis and Saul, 2000), the proposed MCE and four other unsupervised MLs for nonlinear DR and we compared their ability to solve dataset nonlinearity. Furthermore, we compared support vector machine (SVM), classical affinity propagation (AP) and the proposed MCAP for their ability to classify samples projected in reduced feature space.

1.2 Computational biology motivations

H2P ML procedures are extensively employed for image processing (Lattner *et al*., 2004) and for other applications, including bioinformatics (Baldi and Brunak, 1998). A recent study by Cannistraci *et al*. (2009; Ravasi *et al*., 2010), which analyzed genomic transcription factor (TF) measurements, uncovered the presence of specific human tissue patterns. Based on nonlinear DR coupled to clustering in bi-dimensional reduced space, the method offered efficient data visualization and discrimination and, more interestingly, achieved high accuracy in the unsupervised classification of 32 human tissues, on the basis of their embryonic origin. Improvements obtained in the analysis of this dataset are shown in the last part of the article, in which several unsupervised H2P ML methods are compared. However, the main aim of this article is to uncover insights and perspectives that in turn generate solutions for real classification problems in medicine. Accordingly, the first topic selected consists in the development of methods for the classification of subjects with neuropathic pain, which is a major issue in translational and clinical medicine (Baron, 2006; Finnerup and Jensen, 2006). Specifically, we deal with peripheral neuropathy that occurs either with or without pain. Interestingly, some of the patients without pain (NP) can, as the disease progresses, develop a pathological variant with pain (*P*). Since current knowledge of molecular disease mechanisms is poor, no single pain measure has sufficient reliability and validity. New integrative strategies for early diagnosis could greatly enhance the timeliness of therapy planning, and interest in discovering reliable classification and prediction methods for pain patients is accordingly considerable (Baron, 2006; Finnerup and Jensen, 2006; Meyer-Rosberg *et al*., 2001).

Cerebrospinal fluid (CSF) is a valuable source of information for biologists and physicians. A recent computational study analyzed a dataset of 2D electrophoresis (2DE) gel images derived from proteomic CSF profiles of peripheral neuropathic patients (Pattini *et al*., 2008). Control (*C*) and Pain (*P*) groups were partially separated (leave-one-out cross-validation accuracy 68.75%) by a nonlinear surface in the space of the first three principal components extracted by PCA. The discriminative characterization found for patients with pain, along with a further reasons, led us to reconsider this dataset ($n=23$) (Pattini *et al*., 2008). The first additional reason was our interest in assessing the efficiency of differing NML as solutions for the nonlinearity revealed in the profile of pain subjects. Particularly, we tested whether it is possible to solve this nonlinearity by projecting the data in a reduced, 2D space. The result was a clear visualization of proximity and separation between controls and pain subjects, and a minimization of the problem faced by the classifier in finding a line of separation in two dimensions. Our second incentive was that we had the opportunity to follow disease progression; neuropathic patients were still under clinical observation, and four NP patients had developed the clinical features of neuropathic pain (P group). The third reason was the enlargement of the dataset sample to its current total of 42 individuals.

2 DATA AND ALGORITHMS

2.1 Dataset descriptions

The proteomic dataset was obtained from 2DE images generated from CSF samples. 2D gel generation was described in the original proteomics study (Conti *et al*., 2005). Each 2DE image was denoised by median modified wiener filter (MMWF) (Cannistraci *et al*., 2009) and spot detected by means of Progenesis PG240 v2006 software (Nonlinear dynamics, Newcastle, UK). Spot calibration—in accordance with protein chemo-physical coordinates (isoelectric point, pI; relative molecular mass, Mr)—enabled the correction of spot location differences between differing gels. Spot volume was estimated by means of its optical density (sum of the spot pixels) normalized as a percentage of total spot optical density in the gel image (Pattini *et al*., 2008). From each image a vector of 2050 proteomic features was obtained by means of a strategy previously developed, described in depth and validated by Pattini *et al*. (2008). This dataset (dataset 1) was reduced from the original 24 to 23 samples and divided into three groups: $C=8$, NP = 8, $P=7$. As suggested by Pattini *et al*. (2008), we excluded the strongly noised 2DE gel image corresponding to sample P7, which had been used in the previous study exclusively as an internal check. The validation phase introduced a new proteomic dataset ($M=19$) which, together with dataset 1, formed dataset 2. The new M samples derived from a neurological study of amyotrophic lateral sclerosis (ALS) patients not affected by neuropathic pain (Conti *et al*., 2008). The total number of subjects analyzed in dataset 2 of the current study is 42. The demographic and clinical features of dataset 2 subjects/patients are shown in Supplementary Table S1. The dataset is provided on the web site indicated in Section 5.3.

The dataset of human tissues (dataset 3) was provided as supplementary material in the original paper (Ravasi *et al*., 2010) and consists of 32 human tissues and two monocyte cell lines. We exclusively considered human tissues, because cell lines had originally been introduced as an internal check. A total 1321 genomic TF measurements were considered.

2.2 Layout of the neuropathic pain study

A flux graph is provided in Supplementary Figure S1. It clarifies the steps of the H2P procedure, which was used for feature reduction and supervised classification of the proteomic samples, as well as for comparison with the unsupervised H2P variants explained at the end of this paragraph. The layout consists of two stages. The nonlinear mapping of the data in 2D reduced space requires the tuning of a free parameter k that occurs in some MLs for nonlinear dimensionality reduction. This parameter can vary between 1 and $n-1$ neighbors, where n is the dataset sample size, and is generally used to infer local and/or global manifold topology. Here, the idea is also to tune this parameter in order to offer DR projections that are more informative for pain discrimination. The best tunings for LLE (Roweis and Saul, 2000), Gaussian kernel-PCA (KPCA) (Shawe-Taylor and Cristianini, 2004), Local Tangent Space Analysis (LTSA) (Zhang and Zha, 2004) and Isomap (Tenenbaum *et al*., 2000) were learned in order to optimize assignment of subjects to the *C* and *P* samples. This assignment, together with the comparison of ML performance

in solving dataset nonlinearity, was accomplished in 'Stage 1'. In the comparison, a further NML was considered, namely Sammon multi-dimensional scaling (S-MDS), also known as Sammon Mapping (Sammon, 1969). A nonlinear MDS that preserves small distances between data points in the reduced space better than classical MDS, S-MDS is a parameter-free NML that accordingly does not require tuning. In addition, the proposed parameter-free NML called MCE was considered, and its algorithm is presented in Section 2.5.

MCE and LLE offered the best dimensionality reduction for linear discrimination of controls C and subjects with pain P, and they were therefore selected for comparison in the reduction of feature space in the second classification step. In particular, on the basis that LLE was tuned to preserve similarities in relation to the presence or absence of pain, it was also tested in combination with SVM (supervised H2P approach) for the classification of pain neuropathic patients.

In 'Stage 2' design for validation, we propose a procedure in which the SVM classifier is applied in the 2D feature space obtained by LLE; the free LLE parameter is fixed to the best value learned in Stage 1. The SVM classifier also requires a training phase to learn the decision rule used for sample supervised classification in the reduced, 2D feature space. The training of the manifold-NML helps to 'learn the similarities' related to the presence or absence of pain between samples in the high-dimensional feature space, and to map the samples enhancing these similarities in a reduced feature space. In contrast, SVM training helps to 'learn a rule for separation and discrimination of the samples' by exploiting the advantage that the similarities between close samples are enhanced in the new reduced feature space. To ensure robustness, the training both in the 'tuning procedure' and in the 'classification procedure' applied leave-one-out cross validation (LOOCV). For the LOOCV procedure, five of the total eight C and four of the total seven P subjects were randomly selected and used as training exemplars. The dataset was ordered in the following way: $C1-C5$ and $P1-P4$ were used as labels of the training data. The same data were used both for training of the NMLs in Stage 1 and for training of the SVM in Stage 2. The remaining samples were randomly labeled $C6-C8$ and $P5-P7$, and used only for validation.

The validation stage was divided into two tasks:

(i) disease course in NP patients ($n=8$) was predicted as pain or no pain state; in addition, the subjects not used for training ($C6$, $C7$, $C8$ and $P5$, $P6$, $P7$) were also classified by SVM.

(ii) the dataset was extended and the new control patients M ($n=19$) were classified as belonging to the pain or no pain state.

As already mentioned, training a supervised classifier with a small number of samples (five controls versus four pain subjects) in order to infer a model is risky, and it could be further argued that the use of SVM for classification in the reduced linearized feature-extracted space is excessive for this purpose. To address these points, we designed a second H2P approach, completely unsupervised, that substitutes the supervised classifier with an algorithm for unsupervised classification (clustering). Statistical evaluation (accuracy, sensitivity, specificity, precision) of the SVM was performed only on the testing samples and excluding the samples used for training. The unsupervised classification (which does not require training samples) was in turn performed on the entire set of samples in the datasets.

2.3 Minimum Curvilinearity

MC principle has its starting point in the consideration that for datasets of reduced size the idea of estimation or inference of manifold topology in the feature space might be misleading due to the small number of samples. We speculate that in this case it might be more congruous to simply speak of estimation of nonlinear sample distances. MC is proposed as a way to estimate nonlinear sample distances by MST without any need for tuning parameters. A different, interesting principle, summarized in the phrase: 'think globally and fit locally', was introduced with LLE. Exploiting local symmetries of linear reconstructions, LLE is able to learn the global structure of nonlinear manifolds (Roweis and Saul, 2000). This procedure, however, costs the introduction of one free parameter for neighbourhood estimation, which can be a point of weakness in unsupervised tasks. A recent study by Boguñá *et al*. (2009) on the navigability of complex networks found that a general property is present in the hidden metric spaces of several artificial and biological networks. This property is dictated by the shape of the 'hidden metric space', which forces the system to form local interactions between subsets of its elements mapped in the 'observable network topology' as different sub-networks of interacting nodes. On the other hand, the hidden space also guides the greedy-routing process that connects nodes located in differing sub-networks. If this theory is adopted in the framework of our study—applied to the sample representation in the hidden feature space—it offers a valid theoretical support for approaches such as 'think globally and fit locally' and or 'MC'. Indeed, in small datasets, MST provides a reasonably accurate map both of the local connection geometry of near and sub-network-related samples (nodes) and of the global connection geometry between samples (nodes) located in separated regions of the multi-dimensional space.

MC suggests the estimation of curvilinear distances between sample data points in small datasets as pair-wise distances over their MST constructed in the feature space. The collection of all these nonlinear pair-wise distances forms a distance matrix—the MC-distance matrix—to be used as an input in algorithms for DR or clustering.

2.4 Minimum Curvilinear affinity propagation

Although classical AP is a powerful algorithm for clustering that works very well for regularly shaped clusters, with elongated or irregular multi-dimensional data it may force division of single clusters into separate ones or it may provide low-clustering results (Leone *et al*., 2007). The innovation we propose to solve these issues in elongated datasets is a clustering algorithm that runs AP (Frey and Dueck, 2007) over the MST of the samples (here represented in 2D-reduced space). We named this algorithm MCAP clustering. More generally, MCAP is able to define clusters that pass messages between the samples that are nodes on the MST obtained in the multi-dimensional feature space. Taken from another perspective, we can say that this algorithm is MST-guided: we estimate the curvilinear distances between the samples distributed in feature space by MST and then, in accordance with a message passing procedure we send messages between sample points following the preferential highway tracked by MST. Details of the algorithm are reported in Section 5.3.

We compare MCAP results with those offered by the classical AP approach (Frey and Dueck, 2007), where the similarities (negative distances) between the samples are computed as negative squared

error distances (Frey and Dueck, 2007). For MCAP, the similarities are the negative values extracted from the MC-distance matrix. The preference parameter (Frey and Dueck, 2007) for both MCAP and AP is tuned to the value that offers two clusters as an algorithm result.

2.5 Minimum Curvilinear embedding

In terms of comparative ML theory, dimensional reduction and clustering are the two cornerstones of unsupervised learning (Ghahramani, 2004). We can therefore imagine an algorithm for DR that is a distance-matrix analog of the MCAP clustering algorithm. This nonlinear dimensionality reduction algorithm was named MCE and uses the MC-distance matrix as an input for the classical MDS or the nonlinear S-MDS. MCE algorithm details are provided in Section 5.4.

MCE can be interpreted as the 'minimum curvilinear' extension of MDS. The fact that MCE is a nonlinear (and curvilinear) extension of MDS represents a point of similarity with Isomap, which in turn is the manifold geodesic extension of MDS. Isomap computes sample geodesic distances over the manifold as shortest paths on the neighborhood graph, as constructed on the bases of the first k Euclidian-distance neighbors, where k is the free parameter to tune. The principal weakness of Isomap is the algorithm instability encountered in embedding of manifolds with local nonlinearity or discontinuity (Balasubramanian and Schwartz 2002). With the low number of samples available for inferring manifold topology as our starting point, we argue that the strategy of global manifold reconstruction used by Isomap might be not congruous for small and irregular datasets.

3 RESULTS AND DISCUSSION

3.1 Data nonlinearity is successfully addressed

Figure 1A shows algorithms' performance in linear discrimination between controls and pain subjects. Performance optimality was estimated by the maximization of a proposed index for tuning evaluation (TE). This index was evaluated: (i) for increasing values of k, which is the free parameter present in the manifold NML; (ii) for increasing values of standard deviation, which is the free parameter present in the kernel of the Gaussian KPCA. MCE and S-MDS were evaluated by the same index, but they do not present free parameter to tune. The TE index is computed by means of LOOCV in order to prevent overfitting. LOOCV is also used to estimate classification accuracy, as reported in Figure 1B. TE is evaluated as an average measure of linear separation obtained by the removal of one sample per LOOCV round and by the subsequent SVM estimation of the margin of linear separation between the remaining samples. Accuracy is estimated as an average of classification successes calculated by including the sample omitted in the LOOCV round, and by evaluating its label (control or pain) by means of the SVM separation line. Details about TE index and accuracy evaluation are provided in Section 5.1. Figure 1C summarizes the best performance for each tested algorithm.

Data nonlinearity between C and P is successfully addressed during tuning (Stage 1), where four out of six NMLs (LLE, MCE, Isomap, S-MDS) attained linear separation with an accuracy of 1 (Fig. 1B, result not represented for S-MDS), and two out of six (KPCA and LTSA) attained linear separation with an accuracy

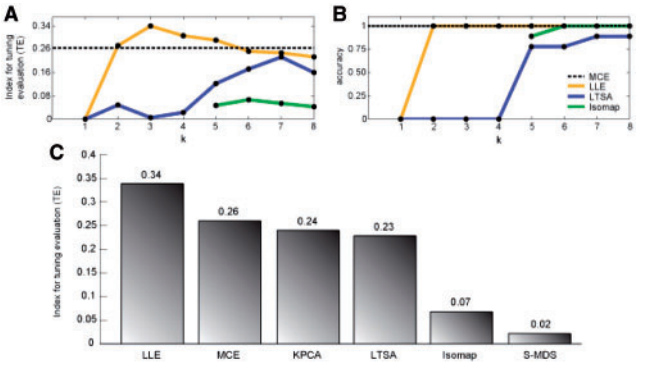

Fig. 1. Tuning and comparison to address the data nonlinearity. (**A**) TE for LLE (yellow line), LTSA (blue line) and Isomap (green line). The x-axis reports different values of neighborhood parameter k; y-axis reports the index for TE. (**B**) Classification accuracy for LLE (yellow line), LTSA (blue line) and Isomap (green line). The x-axis reports different values of neighborhood parameter k; y-axis reports values of accuracy. (**C**) Best performance in linear discrimination compared between tested algorithms.

of 0.89 (eight successes in nine LOOCV rounds; the result for KPCA is not displayed in the figure). This demonstrates that linearization is obtained as a result of generalized NML capacity, and that it is not related to an ability of a single algorithm. In particular, LLE and MCE achieved the highest TE value and they scored the best performance in linear discrimination (Fig. 1A and C). Surprisingly, MCE attained this result without tuning of any parameter and with a score of 1 on accuracy, while LLE was the only manifold NML that enabled high-linear separation for low numbers of neighbors—in a range between 1 and 4—where Isomap and LTSA failed (Fig. 1A and B). Isomap for small k values was not able to recover the manifold structure because the reconstruction of the neighborhood graph was not complete, and this failure did not permit the embedding of the overall number of samples. This weakness of Isomap is due to its topological instability (Balasubramanian and Schwartz 2002). Isomap may construct erroneous connections in the neighborhood graph, and such short-circuits impair its performance (Balasubramanian and Schwartz 2002). The fact that LTSA showed very low performance in the same range where Isomap showed topological instability is a further confirmation of local nonlinearity present in the dataset structure. The locality of the problem is demonstrated by the fact that both algorithms showed inefficiency using a small number of neighbors for manifold reconstruction, and this confirms the result obtained in the previous computational study (Pattini *et al*., 2008). The reasons why Isomap and LTSA show the same behavior, in contrast to LLE, which yields the best performances in this range, are to be found in the differing hypotheses underlying ML applicability. Both Isomap and LTSA are sensitive to the assumption of local manifold linearity (Zhang and Zha, 2004)—which seems to be not satisfied by the dataset considered in our study—whereas LLE provides a local reconstruction that is less sensitive to this assumption. LLE preserves the local properties of the manifold by means of a 'reconstruction weights operation', which locally linearizes—by solving a constrained least-squares problem—the manifold in the neighborhood of each sample (Roweis and Saul, 2000). For datasets with high-intrinsic dimensionality and low number of samples,

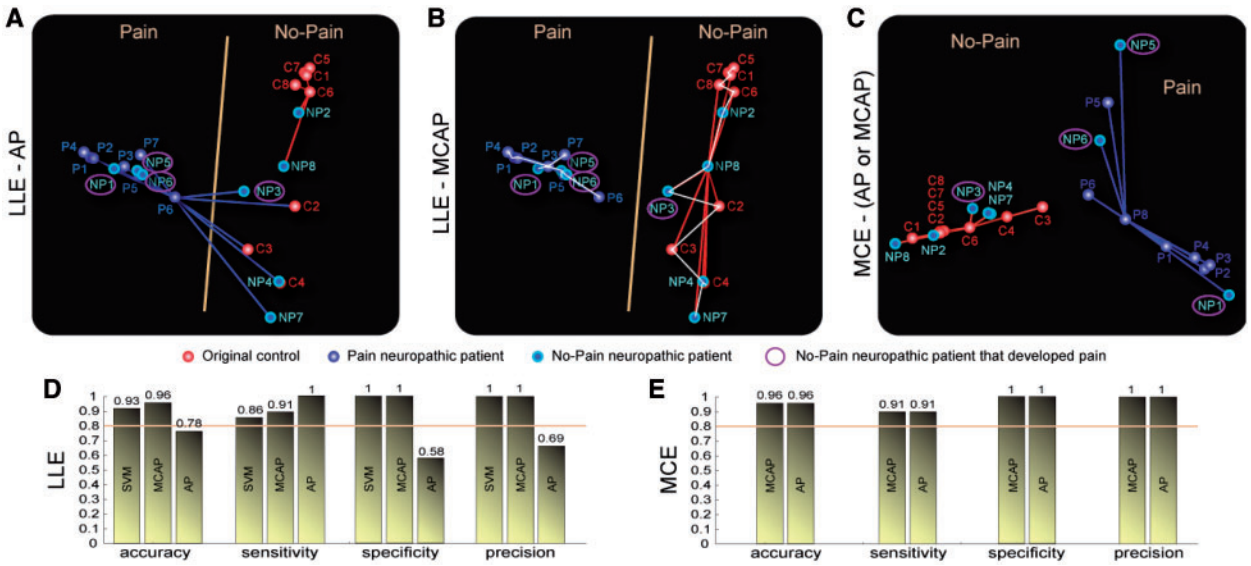

Fig. 2. Evaluation on dataset 1. (**A**) Result of AP clustering in the space of first two dimensions extracted by LLE. (**B**) Result of MCAP clustering in the space of first two dimensions extracted by LLE. (**C**) Results of AP and MCAP clustering in the space of first two dimensions extracted by MCE. Blue line for pain cluster, red line for control cluster. White skeleton in panel (B) indicates the partition generated by MCAP over the MST; link deleted by clustering was between samples *P*6 and NP3. SVM decision rule for classification (yellow line in panels A and B) is obtained considering the training samples ($C1-C5$; $P1-P4$). (**D**) Evaluation of SVM, MCAP and AP for pain classification in LLE reduced space. (**E**) Evaluation of MCAP and AP for pain classification in MCE reduced space.

the underlying estimation of the manifold might be difficult and highly variable. Moreover, the local linearity assumption around certain data points may be violated, at least anisotropically (i.e. only in some manifold directions). Thus, techniques such as Isomap and LTSA may be less successful. In contrast, MCE—designed to estimate nonlinear distances and to deal with local irregularity in small datasets—addresses nonlinearity by considering even the total groups of control (*C*, red spots) and pain (*P*, blue spots) patients present in dataset 1 (Fig. 2C). LLE, as expected, attained comparable performance in this task as well (Fig. 2A and B).

3.2 Prediction and classification of pain subjects

Figure 2 displays the results of different H2P procedures on dataset 1 obtained by combining: (i) LLE with SVM, MCAP or AP; (ii) MCE with MCAP or AP. Although LLE offered a clear linear separation over the first reduced dimension between pain and no-pain subjects, the elongated shape in the bi-dimensional space of the no-pain group caused the failure of AP to identify the right cluster attributions (Fig. 2A). This evidence was already reported in the literature (Leone *et al.*, 2007). In contrast, MCAP succeeded in this task (Fig. 2B) because the message passing procedure was guided by the MST skeleton (white skeleton, Fig. 2B). MCE too provided linear separation over the first reduced dimension (Fig. 2C), but its embedded groups were more regular than those of LLE (Fig. 2A and B), this is why both AP and MCAP provided the same clustering in the MCE reduced space (Fig. 2C). The statistical evaluation displayed in Figure 2D and E suggests that MCE–MCAP, which provided the same result as LLE–MCAP but without any tuning, enjoys high efficiency: a completely unbiased achievement.

This superiority was particularly evident in the second evaluation, performed on dataset 2 (Fig. 3), in which the introduction of the novel sample set *M* caused LLE to shift the linear separation between pain and no-pain state from the first to the second dimension, while the first dimension became discriminative for the various pathological states (Fig. 3A and B). Surprisingly, MCE was still able to discriminate the mixture of five different states over the first dimension (Fig. 3C, data and code to reproduce the figure are provided at the link indicated in Section 5.3): on the left patients affected by ALS neuropathy (*M*); in the centre controls (*C*), while at the bottom-centered peripheral neuropathic patients without pain (NP); on the right, patients with peripheral neuropathy and pain (*P*, and NP with pain). The fact that MCE only needs the first dimension to offer a gradual and shaded landscape of this intricate scenario is impressive, especially if we consider the simplicity of the principle behind this NML, and the absence of parameters to tune. The result of the statistical evaluation displayed in Figure 3D and E suggests that MCE–MCAP provides superior unsupervised discrimination of pain and no-pain subjects, which in turn shows that pain is the prevalent discrimination factor over the first MCE dimension. On the other hand, the performance of the supervised H2P procedure consisting in LLE–SVM (Fig. 3A, B and D) proved to be robust despite the introduction of new samples. However appraisal of this last finding should be tempered by the fact that there were very few test samples.

From the biological point of view, our findings strongly support the efforts to discover reliable methods for the classification of subjects with pain, and encourages speculation about possible ways to distinguish the patients' states in relation to the proteomic pain pattern hidden in their CSF. In order to advance any serious biological claim, a further study with a larger dataset and a congruous investigation of the relation between the significant features is mandatory, yet this result is important because of

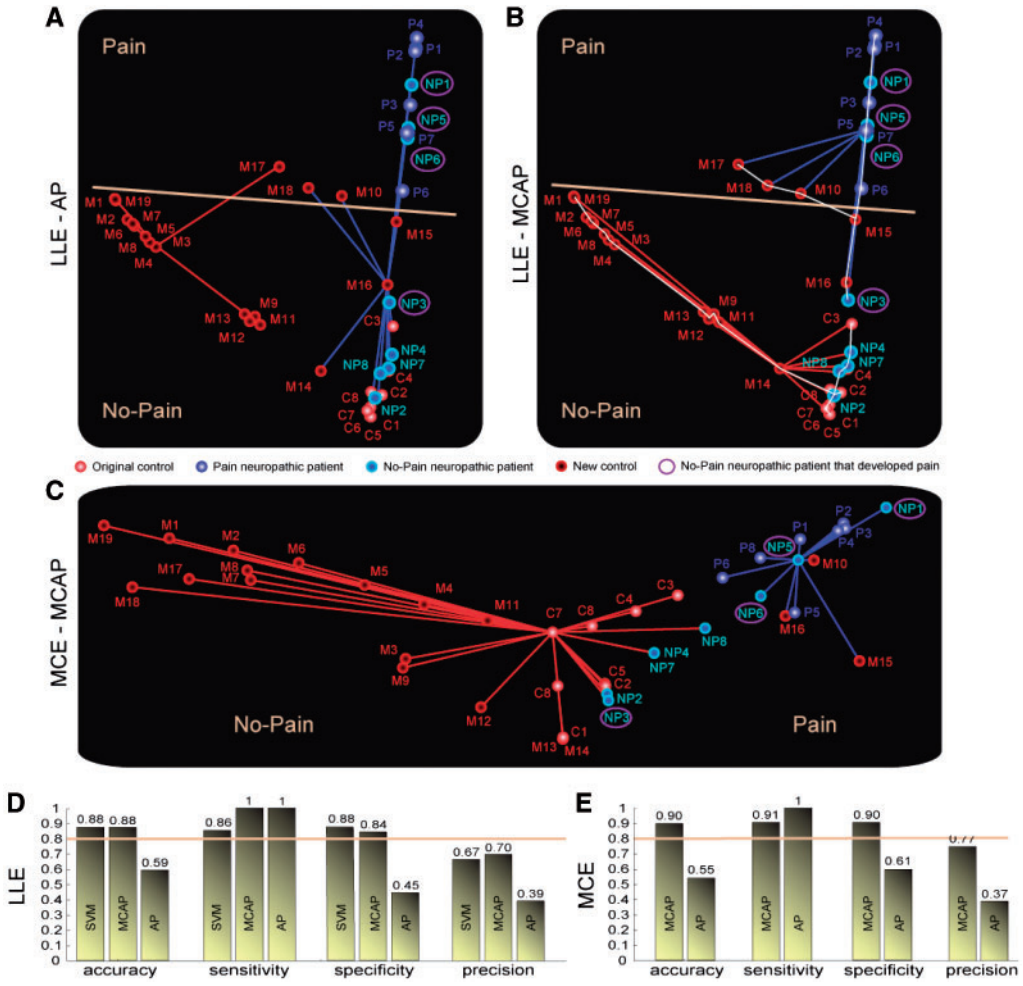

Fig. 3. Evaluation on dataset 2. (**A**) Result of AP clustering in the space of first two dimensions extracted by LLE. (**B**) Result of MCAP clustering in the space of first two dimensions extracted by LLE. (**C**) Results of MCAP clustering in the space of first two dimensions extracted by MCE. Blue line for pain cluster, red line for control cluster. White skeleton in panel (B) indicates the partition generated by MCAP over the MST; link deleted by clustering was between samples *C*3 and NP3. SVM decision rule for classification (grey line in panel A and B) is obtained considering the training samples ($C1-C5$; $P1-P4$). (**D**) Evaluation of SVM, MCAP and AP for pain classification in LLE reduced space. (**E**) Evaluation of MCAP and AP for pain classification in MCE reduced space.

the high-throughput proteomic screening approach employed to characterize every sample. An interesting final note from the clinical standpoint is that the unsupervised analysis provided accurate identification (only one misclassification NP3, out of a total eight NP samples) of future pain for NP patients (Figs 2A–C and 3A–C). This result is summarized in Supplementary Table S2 (see the 'computationally predicted state' column), together with the clinical follow-up at 6–12 months and at >1 year (see 'follow up' columns).

3.3 The tissue embryological classification is improved

On dataset 3, MCE and LLE (Fig. 4A and B) demonstrated the best dimensionality reduction (same clustering accuracy, Fig. 4D) by solving the nonlinearity better than Gaussian KPCA (Fig. 4D), while PCA performance was much lower (Fig. 4C and D). Isomap suffered from instability and its DR was not effective for evaluation. The ability of MCE to provide a discriminative landscape where the sample classes are gradually unfolded along the first dimension is maintained in this dataset too (Fig. 4A). In contrast, and as in the previous evaluation, LLE, needs to combine the first and second dimensions for a complete discrimination of the classes (Fig. 4B). As previously mentioned, this result in DR by MCE is very important because it is completely unbiased (absence of free parameter to tune in the algorithm). LLE allowed best clustering considering $k=5$ both for MCAP and AP, and this value was tuned on the basis of knowledge of the sample labels. Surprisingly, the results for MCE and LLE are not only similar in accuracy, but also in cluster shape and in the co-localization of differing samples, especially in the endodermal cluster (red color, Fig. 4A and B). We do not have any biological explanation for the misclassification of lymph-node (22, Fig. 4A and B), but it was suggested (Dorshkind, 2002) that bone marrow (21, Fig. 4A and B) might also contain distinct endodermal progenitors capable of contributing to components of

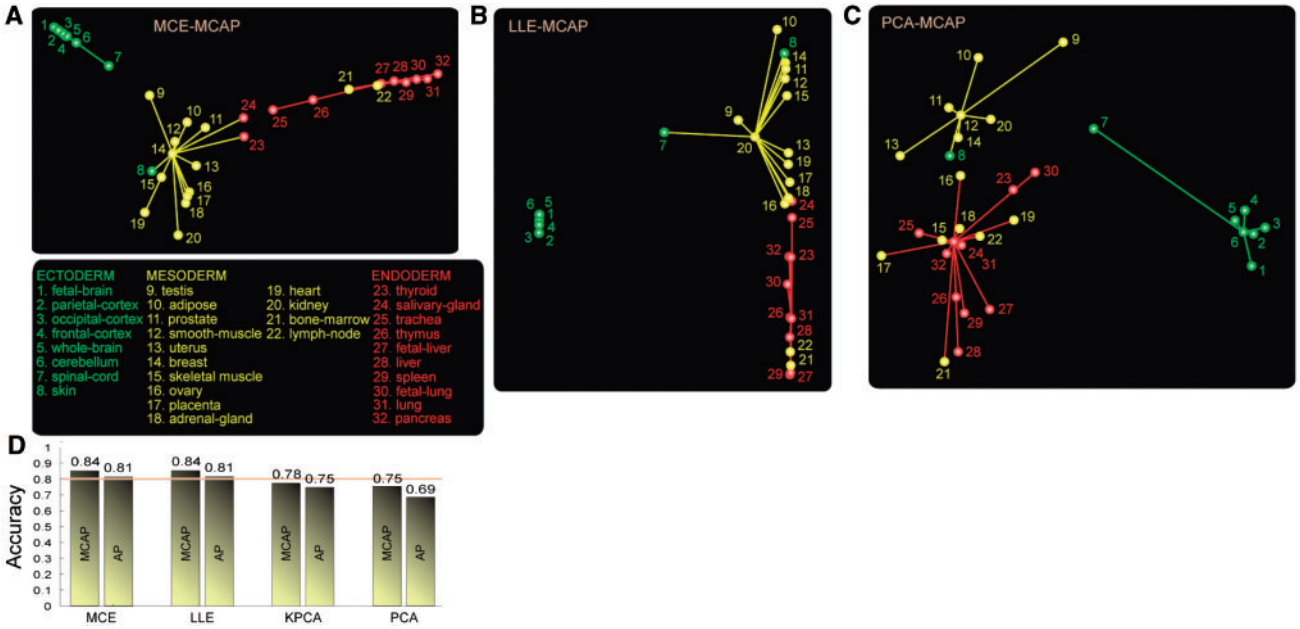

Fig. 4. Evaluation on dataset 3. (**A**) Result of MCAP clustering in the space of first two dimensions extracted by MCE. (**B**) Result of MCAP clustering in the space of first two dimensions extracted by LLE. (**C**) Results of MCAP clustering in the space of first two dimensions extracted by PCA. Green line for ectodermal cluster attribution, yellow line for mesodermal cluster attribution, red line for endodermal cluster attribution. (**D**) Evaluation of MCAP and AP for unsupervised classification in reduced space obtained by different methods.

the gastrointestinal system such as liver (27 and 28, Fig. 4A and B). Conversely, for MCE the mesodermal attribution of thyroid and salivary-gland samples (23 and 24, Fig. 4A) seems to be an error due to limitations of unsupervised classification rather than to a misleadingly low-dimensional localization. MCAP invariably offered better accuracy (Fig. 4D) than did AP, and thus confirmed the results previously obtained on datasets 1 and 2. The most impressive result is that the MCE–MCAP H2P procedure achieved 84% accuracy on the genomic TF expressions in a completely unsupervised manner. This is an improvement that supports the result (82% accuracy) reported in the previous article (Ravasi *et al*., 2010) by the small (only six interactions) TF-homeobox subnetwork, and confirms both the procedure's power in embryological discrimination and its potential importance in tissue differentiation processes.

Interestingly, skin (labeled as ectodermal) was always classified in the mesodermal cluster in each of the different ML analyses (label 8, Fig. 4A–C). This is in accordance with the classification attained in the original article (Ravasi *et al*., 2010), but there interpreted as misclassification. In the light of the latest results, a possible mesodermal re-attribution of the skin label could be considered. The biological explanation resides in the multi-layer structure of the skin: although the first layer of the skin (epidermis) is ectodermal, the extracted skin sample might also contain the second layer (dermis), which is of mesodermal origin.

4 CONCLUSION AND FUTURE PERSPECTIVES

MCE and LLE were very effective for DR because they allowed similarly high-clustering accuracy, and they occasionally uncovered analogous geometry in sample localization (Fig. 4A and B). We speculate that these similarities between MCE and LLE results are evidence of closeness—as far as small datasets are concerned—between the principles of 'MC' and 'think globally and fit locally' that respectively underlay MCE and LLE. Moreover, the fact that MCE only required the first dimension to completely unfold as many as five different classes (Fig. 3C) is striking, especially if we consider the simplicity of the principle this NML is based upon, and the absence of a parameter to tune. In our evaluations, LLE needed the first and second dimensions to yield the same discriminative results, and this comparable performance was obtained at the cost of a free parameter to tune. If no label hypothesis is provided (as was the case for tissue embryological attribution, the uncovering of which was unsupervised), it is hard to imagine an unsupervised strategy that indicates the right tuning for unfolding the classes hidden in a nonlinear dataset. However, we showed that this defect can be transformed into a merit through combination with supervised classifiers like SVM, where the tuning parameter—such as a kernel parameter—can be used to enhance sample discrimination in relation to the aim of the supervised task.

We expect PCA to exceed MCE on linear data, and for practical applications we accordingly suggest initial use of PCA in combination with differing normalizations; if the dataset shows subsequent resistance and nonlinearity (as in Fig. 4C), we recommend structural exploration by means of MCE (as in Fig. 4A) and other NML techniques. Another solution could be the direct employment of the MCE–MCAP H2P approach, which in our results proved to be very powerful for visualization and unsupervised classification. In particular, MCAP overcame AP in the clustering of elongated data in the bi-dimensional reduced space, but we expect that, for regularly shaped clusters, AP might perform similarly or better.

Although the MC principle provided a valuable framework for the estimation of curvilinear distances in small, nonlinear, multi-dimensional datasets, its extension to large datasets needs careful consideration and adaptation. In such extension, the MST-measure may not be sufficient to estimate the distances over the manifold with adequate approximation. Large numbers of samples can cause the overestimation of sample distances over the MST, and these large distances could prevail in magnitude over the shorter ones in the low-dimensional representation. An idea for future developments is the extension of the MC approach to other ML algorithms. In this perspective, an option for future investigation might consider the minimum curvilinear LLE (MCLLE), in which neighbors are estimated by distances over the MST and not, as in classical LLE, by Euclidean distances (EDs). On the other hand, there might be additional benefit in the use of re-sampling techniques to compute more refined estimations of manifold topology, where one possible solution would be to estimate pair-wise distances by bootstrapping samples and/or features.

To the best of our knowledge, the current study is the first to derive unsupervised classification of pain onset from CSF proteomic profiles, and this result could offer new insights for the future characterization of pain in molecular and systems biology.

A final observation regards tissue embryological classification. Prompted by the improved accuracy here reported, we suggest that the skin label might be reinterpreted as a mesodermal attribution. We also conjecture that data nonlinearity could be completely addressed by methods for DR that more finely exploit the intrinsic patterns hidden in biological TF-network topology.

5 METHODS

5.1 Tuning stage: for addressing data nonlinearity

Index TE is estimated by means of LOOCV in order to prevent overfitting. Having fixed the value of the free parameter k, the considered DR algorithm provides a two-DR for each leave-one-out step, excluding one sample from the training dataset in each round of LOOCV (nine samples in the dataset provide nine DRs during the leave-one-out procedure), and then estimating a proposed cluster validity measure (CVM) (Stein *et al*., 2003) in the 2D reduced space. The CVM is here used as a measure of separation between the two classes C and P present in the training dataset. Higher CVM values mean better separation between the two considered groups in the 2D-reduced space; in the absence of linear separation CVM provides value zero. Thus, for every value of k an ensemble of CVMs is computed during LOOCV. This ensemble is adopted to calculate the index TE in correspondence to any value of the free parameter k according to the following formula:

$$\mathrm{TE}(k) = \frac{\mathrm{mean}(\mathrm{CVMs})}{1+\mathrm{SD}(\mathrm{CVMs})}.$$

The mean is divided by the SD (+1) of the CVMs. This estimation is used to measure the training optimality of the considered algorithm in correspondence to each value assumed by the free parameter k. For 'SD' equal to zero, TE takes a value corresponding to the mean CVM. For 'SD' >0, TE is penalized with respect to the mean CVM. In the absence of linear separation, TE has zero value if the CVM has zero value for each of the LOOCV steps. The rationale is to select the parameter value that offers high-cluster separation (high-CVM mean value) and at the same time ensures high reliability and robustness (low-CVM SD) during the cross-validation procedure. Details on CVM and accuracy computing, as well as details of the toolbox used for implementation of LLE, LTSA and Isomap algorithms, are provided in Supplementary Data (paragraph 1).

5.2 Validation Stage 2: prediction and classification of pain subjects

During the 'validation Stage 2', DR—in a 2D space—of the entire dataset was obtained by exploiting the best parameter setting $k=3$, which was learned for the LLE algorithm during the 'tuning stage'. An ensemble of decision boundary (DB) was subsequently obtained, by means of SVM and the procedure based on the LOOCV (described above), using the same C ($n=5$, $C1$, …, $C5$) and P ($n=4$, $P1$, …, $P4$) samples as those previously employed in the training of the 'tuning stage'. The DB offering median distance between the support vectors was designated as the decision rule.

5.3 MCAP

The first step is to calculate a distance matrix (MC-matrix) as pair-wise sample distances over the MST, as computed by the Kruskal method in the feature space (in our case, 2D-reduced space). For MST computation, we suggest the use of a heuristic metric that we found fitted efficiently in combination with the message passing procedure run by AP over the MST. The suggested heuristic is the square root of the EDs between the samples. This device attenuates the estimation of large distances and amplifies the estimation of short distances; consequently the device helps to regularize the distances over the MST for the message passing procedure. In the second step, AP is run assuming sample similarities equal to the negative values of the elements in the MC-matrix—computed as previously described—and tuning the preference parameter (Frey and Dueck, 2007) to the value that offers two clusters as algorithm result. Matlab code at: https://sites.google.com/site/carlovittoriocannistraci/home http://www.mathworks.com/matlabcentral/ (tag: MC).

5.4 MCE

The first step is to calculate a distance matrix (MC-matrix) as pair-wise sample distances over the MST as computed by the Kruskal method in the feature space. To compute the MST in the feature space, we tested the ED and the correlation distance (CD) obtained as:

$$\mathrm{corr}(x,y) = 1 - \mathrm{corr}_{\mathrm{person}}(x,y)$$

In general, the two different distances provided comparable results. We used the CD in our computation, except for the analysis of dataset 2, in which the ED was preferred. In the second step we performed the embedding transformation by performing the classical MDS of the MC-matrix. We also tested Sammon nonlinear MDS (S-MDS), but the result on our data, although comparable, was less impressive. Matlab code is provided on the web sites indicated in Section 5.3.

ACKNOWLEDGEMENTS

We thank Ewa Aurelia Miendlarzewska for her generous assistance and for language revision and Sven Bergmann for precious suggestions.

Funding: Fondazione CARIPLO (NOBEL GuARD Project); MoH RF-FSR-2007-637144; Italian Interpolytechnic School of Doctorate SIPD (http://sipd.polito.it/) (to C.V.C.); US National Institute of Mental Health and the King Abdullah University of Science and Technology (grant MH062261 to T.R. and C.V.C.).

Conflict of Interest: none declared.

REFERENCES

Balasubramanian,M. and Schwartz,E.L. (2002) The Isomap algorithm and topological stability. *Science*, **295**, 7.

Baldi,P. and Brunak,S. (eds) (1998) Hybrid systems: hidden Markov models and neural networks. In *Bioinformatics: The Machine Learning Approach*. The MIT Press, Cambridge, MA, USA.

Baron,R. (2006) Mechanisms of disease: neuropathic pain–a clinical perspective. *Nat. Clin. Pract. Neurol.*, **2**, 95–106.

Boguñá,M. *et al*. (2009) Navigability of complex networks. *Nat. Phys.*, **5**, 74–80.

Cannistraci,C.V. *et al*. (2009) Median-modified Wiener filter provides efficient denoising, preserving spot edge and morphology in 2-DE image processing. *Proteomics*, **9**, 4908–4919.

Conti,A. *et al*. (2005) Pigment epithelium-derived factor is differentially expressed in peripheral neuropathies. *Proteomics*, **5**, 4558–4567.

Conti,A. *et al*. (2008) Differential expression of ceruloplasmin isoforms in the cerebrospinal fluid of Amyotrophic Lateral Sclerosis patients. *Proteomics Clin. Appl.*, **2**, 1628–1637.

Dorshkind,K. (2002) Multilineage development from adult bone marrow cells. *Nat. Immunol.*, **3**, 311–313.

Finnerup,N.B. and Jensen,T.S. (2006) Mechanisms of disease: mechanism-based classification of neuropathic pain-a critical analysis. *Nat. Clin. Pract. Neurol.*, **2**, 107–115.

Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Ghahramani,Z. (2004) Unsupervised learning. In Bousquet,O. *et al*. (eds) *Advanced Lectures on Machine Learning*. Springer, Berlin, Germany, pp. 72–112.

Lattner,A.D. *et al*. (2004) A combination of machine learning and image processing technologies for the classification of image regions. In *Proceedings of the Adaptive Multimedia Retrieval (Lecture Notes in Computer Science, LNCS series)*. Springer, Berlin/Heidelberg, pp. 341–351.

Leone,M. *et al*. (2007) Clustering by soft-constraint affinity propagation: applications to gene-expression data. *Bioinformatics*, **23**, 2708–2715.

Martella,F. (2006) Classification of microarray data with factor mixture models. *Bioinformatics*, **22**, 202–208.

Meyer-Rosberg,K. *et al*. (2001) Peripheral neuropathic pain - a multidimensional burden for patients. *Eur. J. Pain*, **5**, 379–389.

Pattini,L. *et al*. (2008) An integrated strategy in two-dimensional electrophoresis analysis able to identify discriminants between different clinical conditions. *Exp. Biol. Med.*, **233**, 483–491.

Ravasi,T. *et al*. (2010) An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell*, **140**, 744–752.

Roweis,S.T. and Saul,L.K. (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

Sammon,J.W. (1969) A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.*, **18**, 401–409.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA.

Smialowski,P. *et al*. (2009) Pitfalls of supervised feature selection. *Bioinformatics*, **26**, 440–443.

Stein,B. *et al*. (2003) On cluster validity and the information need of users. In Hanza,M.H. (ed.) *Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 03)*. ACTA Press, Benalmadena, Spain, pp. 216–221.

Tenenbaum,J.B. *et al*. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.

Zhang,Z. and Zha,H. (2004) Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM J. Sci. Comput.*, **26**, 313–338.