SUPPLEMENTARY INFORMATIONS

1. Tuning Stage: for addressing data nonlinearity

The CVM is based on the separation line obtained in the 2D-reduced space by linear-kernel hard-margin SVM which is a maximal margin classifier (Cristianini and Shawe-Taylor, 2000). The separation line could be interpreted also as a decision boundary (DB) between the two groups C and P considered in the classification. The CVM is defined as in the following formula:

$$CVM = \frac{MM}{1 + s(C_1) + s(C_2)}$$

MM is an inter-cluster measure constructed as the distance between the opposite support vectors (the maximum margin identified by SVM) of the two clusters linearly separated by SVM. The term s(C_i) indicates a scatter measure of the cluster $C_{i.,}$ thus an intra-cluster measure. We adopted the following formula:

$$s(C_i) = \exp\{\max[\Delta(c_{i,}, p_i)]^2\} - 1$$

The maximum deviations Δ - measured as the Euclidean distance between the cluster centre c_i and the generic cluster points p_i – is estimated by means of a Gaussian function exp(x^2) in order to penalize clusters that present outlier points or that are excessively spread. In absence of linear separation CVM outcome is zero, MM being equal to zero. The SVM classifier was implemented in MATLAB using the function 'quadprog' as described in (Cristianini and Shawe-Taylor, 2000).

Classification accuracy is also evaluated in correspondence to any value assigned to parameter *k*, according to the LOOCV procedure used for estimating the TE. In each round of the LOOCV one sample is excluded from the training dataset and then the correctness of its classification is verified with respect to the SVM separation line (previously computed in 2D-reduced space for CVM estimation). The measure of the predictive accuracy is obtained as the percentage of successes.

We adopted LLE, LTSA and Isomap codes as implemented in the version 0.7b of the Matlab Toolbox for Dimensionality Reduction (Laurens Van der Maaten, Maastricht University, 2007), freely available at:
http://www.cs.unimaas.nl/l.vandermaaten/Laurens_van_der_Maaten/Matlab_Toolbox_for_Dimensionality_Reduction.html.
For more details refer to the toolbox user guide (L.J.P. Van der Maaten - An Introduction to Dimensionality Reduction Using Matlab - Report MICC 07-07. Maastricht University, The Netherlands). Computation was implemented using MATLAB v. 7.0 (The MathWorks^{TM}).

Table 1

#	Gender	Age	Diagnosis	CSF protein [C] (mg/ml)
C1	M	62	Headache	0.55
C2	F	30	Minor ortopedic surgery	0.23
C3	F	30	Minor ortopedic surgery	0.30
C4	F	43	Minor ortopedic surgery	0.30
C5	F	50	Minor ortopedic surgery	0.13
C6	M	34	Minor ortopedic surgery	0.36
C7	M	56	Headache	0.23
C8	F	25	Idiopathic intracranial hypertension	0.20
NP1	M	50	IgM anti MAG	1.06
NP2	M	73	Toxic	0.81
NP3	M	38	Toxic	0.47
NP4	M	60	Idiopathic	0.33
NP5	F	64	Toxic	0.47
NP6	M	74	Paraneoplastic	0.32
NP7	F	42	Idiopathic	0.53
NP8	M	64	Idiopathic	0.25
P1	F	50	Inflammatory	0.23
P2	M	31	Inflammatory	0.77
P3	M	35	Toxic	0.51
P4	F	53	Inflammatory	0.57
P5	M	47	Legs mononeuropathy	0.22
P6	F	60	Inflammatory	0.35
P8	M	46	Idiopathic	0.45
MND1	F	50	SMA	0.27
MND2	M	69	sALS	0.11
MND3	M	22	SMA	0.41
MND4	M	50	sALS	0.54
MND5	F	68	sALS	0.30
MND6	M	61	sALS	0.27
MND7	F	70	sALS	0.46
MND8	F	53	SMA	0.25
MND9	M	63	SMA	0.28
MND10	M	57	SMA	0.58
MND11	M	50	sALS	0.80
MND12	F	80	SMA	0.48
MND13	M	56	sALS	0.25
MND14	M	32	SMA	0.25
MND15	F	30	sALS	0.33
MND16	M	52	sALS	0.55
MND17	M	50	sALS	0.63
MND18	M	64	sALS	0.30
MND19	M	54	sALS	0.34

C= control haelthy subject; NP= peripheral neuropathy without pain;
P= peripheral neuropathy with pain; MND= motor neuron disease;
sALS= sporadic amyotrophic lateral sclerosis; SMA= spinal muscular atrophy

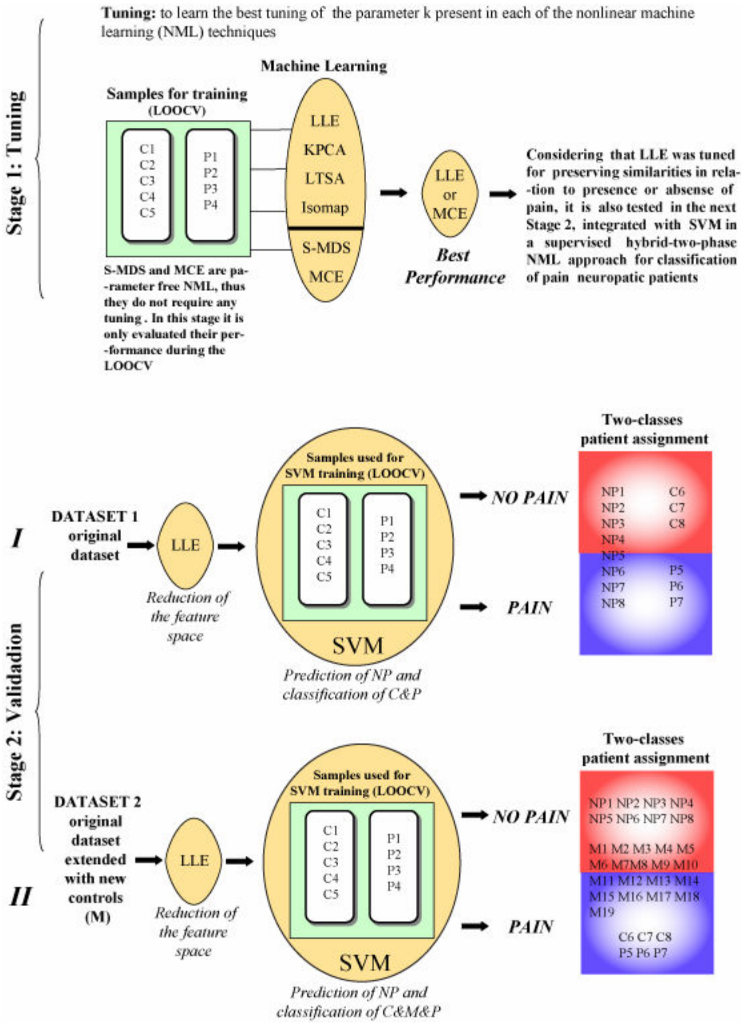
Stage 1: Tuning
Tuning: to learn the best tuning of the parameter k present in each of the nonlinear machine learning (NML) techniques
Machine Learning
Samples for training (LOOCV)
C1 C2 C3 C4 C5
P1 P2 P3 P4
LLE
KPCA
LTSA
Isomap
S-MDS
MCE
LLE or MCE
Best Performance
Considering that LLE was tuned for preserving similarities in rela--tion to presence or absense of pain, it is also tested in the next Stage 2, integrated with SVM in a supervised hybrid-two-phase NML approach for classification of pain neuropatic patients
S-MDS and MCE are pa--rameter free NML, thus they do not require any tuning . In this stage it is only evaluated their per--formance during the LOOCV
Stage 2: Validadion
I
DATASET 1 original dataset
LLE
Reduction of the feature space
Samples used for SVM training (LOOCV)
C1 C2 C3 C4 C5
P1 P2 P3 P4
SVM
Prediction of NP and classification of C&P
NO PAIN
PAIN
Two-classes patient assignment
NP1 NP2 NP3 NP4 NP5 NP6 NP7 NP8
C6 C7 C8
P5 P6 P7
II
DATASET 2 original dataset extended with new controls (M)
LLE
Reduction of the feature space
Samples used for SVM training (LOOCV)
C1 C2 C3 C4 C5
P1 P2 P3 P4
SVM
Prediction of NP and classification of C&M&P
NO PAIN
PAIN
Two-classes patient assignment
NP1 NP2 NP3 NP4 NP5 NP6 NP7 NP8
M1 M2 M3 M4 M5 M6 M7M8 M9 M10
M11 M12 M13 M14 M15 M16 M17 M18 M19
C6 C7 C8 P5 P6 P7

Fig. 1. Layout of the study. C, healthy control subjects; P, peripheral neuropathy patients with pain; NP, peripheral neuropathy patients without pain; M, subjects with motor neuron disease without pain employed as controls for pain classification. NML, nonlinear machine learning; LLE, Locally Linear Embedding; KPCA, Gaussian Kernel Principal Component Analysis; LTSA, Local Tangent Space Analysis; S-MDS, Sammon Multidimensional Scaling; MCE, Minimum Curvilinear Embedding; SVM, Support Vector Machine. LOOCV, leave-one-out cross validation.

In classification, a reduction of the feature space is recommended. It decreases the complexity and improves performance (Kohavi and Rand, 1997; Lai, et al., 2006). Techniques for dimensionality reduction are preferred to feature selection for analysis of datasets of reduced sample dimension, because in general they are unsupervised algorithms (Varshavsky, et al., 2006). Feature selection, which is usually supervised, requires a larger number of samples to be correctly applied, being prone to over-training for a small number of samples (Lai, et al., 2006; Smialowski, et al., ; Varshavsky, et al., 2006). In 2DE computational proteomics the use of dimensionality reduction is frequent because the number of samples is limited to several tens for technical and experimental reasons (Gottfries, et al., 2004; Marengo, et al., 2005; Marengo, et al., 2008; Marengo, et al., 2006; Marengo, et al., 2003; Pattini, et al., 2008). These reasons motivate our choice to use dimensionality reduction not only for data visualization, but also for solving the nonlinearity related to the proteomic profile of the pain patients, extracting meaningful meta-features for subsequent classification.

Table 2. Characteristics of neuropathic patients without pain (NP)

Patient	Neuropathy Diagnosis	Initial Neuropathic State	Computationally Predicted State	Follow-up (6-12 months)	Follow-up (>12 months)
NP1	IgM anti MAG	No Pain	Pain	Pain (6 mm)	Pain
NP2	Toxic	No Pain	No Pain	No Pain	No Pain
NP3	Toxic	No Pain	No Pain	No Pain	Pain (13 mm)
NP4	Idiopathic	No Pain	No Pain	No Pain	No Pain
NP5	Toxic	No Pain	Pain	No Pain	Pain (15 mm)
NP6*	Paraneoplastic	No Pain/Pain*	Pain	No Pain/Pain*	Pain*
NP7	Idiopathic	No Pain	No Pain	No Pain	No Pain
NP8	Idiopathic	No Pain	No Pain	No Pain	No Pain

* The NP6 patient is affected by peripheral neuropathic pain due to nerve injury by hernia compression in lumbar radiculopathy, thus not related to the original diagnosis of paraneoplastic neuropathy

An interesting point, from the clinical standpoint, regards the NP6 patient who was identified as a potential pain patient by our approach. This patient, suffering from paraneoplastic neuropathy, did not develop neuropathic pain due to this form of PN during the disease progression. However, from the beginning of the hospitalization he showed strong pain (in a different area: right leg) related to a lumbar radiculopathy generated by hernia compression. This prediction is interesting because, although it is not related to the paraneoplastic PN clinical progression, the radiculopathy is one of the different PNs which causes peripheral neuropathic pain (Meyer-Rosberg, et al., 2001). In fact, compression of the nerve root due to hernia is a nerve injury which generates inflammatory state coupled to pain in the patient (Meyer-Rosberg, et al., 2001).

References

Cristianini, N. and Shawe-Taylor, J. (2000) *An introduction to Support Vector Machine and other kernel-based learning methods*. Cambridge University Press.

Gottfries, J., Sjogren, M., Holmberg, B., Rosengren, L., Davidsson, P. and Blennow, K. (2004) Proteomics for drug target discovery *Chemometrics and Intelligent Laboratory Systems*, 73, 47-53.

Kohavi, G. and Rand, J. (1997) Wrappers for Feature Subset Selection, *Artificial Intelligence*, 97, 273-324.

Lai, C., Reinders, M.J., van't Veer, L.J. and Wessels, L.F. (2006) A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets, *BMC Bioinformatics*, 7, 235.

Marengo, E., Robotti, E., Antonucci, F., Cecconi, D., Campostrini, N. and Righetti, P.G. (2005) Numerical approaches for quantitative analysis of two-dimensional maps: a review of commercial software and home-made systems, *Proteomics*, 5, 654-666.

Marengo, E., Robotti, E., Bobba, M., Demartini, M. and Righetti, P.G. (2008) A new method of comparing 2D-PAGE maps based on the computation of Zernike moments and multivariate statistical tools, *Anal Bioanal Chem*, 391, 1163-1173.

Marengo, E., Robotti, E., Bobba, M., Liparota, M.C., Rustichelli, C., Zamo, A., Chilosi, M. and Righetti, P.G. (2006) Multivariate statistical tools applied to the characterization of the proteomic profiles of two human lymphoma cell lines by two-dimensional gel electrophoresis, *Electrophoresis*, 27, 484-494.

Marengo, E., Robotti, E., Gianotti, V., Righetti, P.G., Cecconi, D. and Domenici, E. (2003) A new integrated statistical approach to the diagnostic use of two-dimensional maps, *Electrophoresis*, 24, 225-236.

Meyer-Rosberg, K., Kvarnström, A., Kinnman, E., Gordh, T., Nordfors, L. and Kristofferson, A. (2001) Peripheral neuropathic pain - a multidimensional burden for patients., *European journal of pain*, 5, 379-389.

Pattini, L., Mazzara, S., Conti, A., Iannaccone, S., Cerutti, S. and Alessio, M. (2008) An integrated strategy in two-dimensional electrophoresis analysis able to identify discriminants between different clinical conditions, *Exp Biol Med (Maywood)*, 233, 483-491.

Smialowski, P., Frishman, D. and Kramer, S. Pitfalls of supervised feature selection, *Bioinformatics*, 26, 440-443.

Varshavsky, R., Gottlieb, A., Linial, M. and Horn, D. (2006) Novel unsupervised feature filtering of biological data, *Bioinformatics*, 22, e507-513.