Dr Peter Tino

Theo Murphy Blue Skies awards - 2009

Title:	Dr
First Name:	Peter
Surname:	Tino
Other Names:	
Honours:	
Address:	School of Computer Science The University of Birmingham Edgbaston
Town:	BIRMINGHAM
Postcode:	B15 2TT
Country:	United Kingdom
Nationality:	**Nationality** Slovak
Email Address:	P.Tino@cs.bham.ac.uk
Web Address:	http://www.cs.bham.ac.uk/~pxt/
Telephone (work):	121 414 8558
Fax:	121 414 4281

Applicant Career Summary

Statement of qualifications and career:	
Publications:	see attached document
Present Position:	Senior Lecturer
Present Department:	School of Computer Science
Present Position Description:	This is a permanent position (since January 2003). I teach 2 modules per year; manage all final year projects; supervise undergrad, MSc and PhD students
Pending Applications:	
Where did you hear of this scheme?:	University sources
Existing grants:	

Proposal

Project Title:	A New Framework for Holistic Representations and Learning on Aminoacid Sequences
Collaboration:	Biological interpretation of the results will be done in close collaboration with Dr. Mikael Boden (and his colleagues) at the Institute of Molecular Bioscience, University of Queensland. I have had an active collaboration with Dr. Boden on processing temporal sequence structures via recurrent neural networks (publication [30]). We will communicate over email and phone/skype.
Research proposal:	see attached document
Need for teaching relief:	I am normally involved in teaching 2 modules. I am also managing all final year projects within the School of Computer Science, supervising 4 PhD students and a number of undergraduate and MSc projects. The proposed is doable within a year, but will require full concentration of resources. That is why I am asking for a full teaching and administration buy-out for a year. If successful, I will continue to supervise my PhD students and sit on the committees (Research Student Monitoring Group and Research Committee).
Comply with Policy on use of Animals:	Not applicable
Ethical permission:	
Lay Report:	

Today's experimental technologies in biology offer biologists access to large volumes of unprecedented data. This technological progress is driving efforts to uncover the function of whole-genome data with a huge potential for personalised health care. However, for such a revolution to happen we need to be capable to explore and learn from such large scale data in an automated manner.

One of the challenges of automated learning on biological data is their highly structured nature.
This proposal will develop a novel way of representing and learning on protein data that takes into account as much biologically useful information as possible. Such biologically rich representations would be problematic using previous approaches.

Proteins (basic building blocks for living organisms) are long molecules consisting of a series of chained aminoacids. There are about 20 different aminoacids and each protein can be viewed as a string over 20 symbols (one symbol per aminoacid). However, protein molecules can form complicated 3-dimensional shapes that, together with physical-chemical properties of aminoacids determine the protein function. How can such structure be represented for automated learning to take place? We base our approach on 4 key assumptions:
1. All potentially important information on proteins should be represented.
2. The learning machine should be lead to focus on the dominant trends in the data set that are crucial for the learning task.
3. As the nature of the task changes, so does the view on what are the important and dominant factors in the structure-rich data.
4. Representing a rich variety of potentially important information on proteins can lead to too many degrees of freedom, hampering the effectiveness of the learning process.

We address all 4 points in a unified framework.
Each protein will be represented as a binary tensor. A tensor is a generalisation of the notion of vector, where the elements are indexed along more than one direction (mode). For example, an order-2 tensor is a matrix with two modes - rows and columns. For proteins we suggest to have e.g. 4 modes - each protein will be represented by a 4th-order tensor. Mode 1 will index position within the aminoacid sequence,
important motifs of aminoacid groups will be represented along mode 2, local spatial structure of the protein molecule along mode 3 and physical-chemical properties of the relevant aminoacids along mode 4.
We then perform a task-driven compression of the data tensors in the tensor space, thus achieving a reduced representation of the dominant trends in the data that are crucial for the given task and enhance the model generalisation.
The framework will be verified on an important and difficult problem of nuclear protein localisation (e.g. the process by which proteins get transported into the cell nucleus).
Understanding nuclear localisation of proteins is crucial for understanding the dynamics and self-regulation of the cell.

This truly interdisciplinary research will bridge machine learning, bioinformatics and biology. If successful, our proposal can make a high impact in the bioinformatics field, but needs a proof of concept before further major developments can take place. Relying on task-driven compression of vast sparse binary tensors representing proteins in a holistic manner is a novel concept without sufficient evidence base and as such would be considered as lacking feasibility component by traditional grant schemes.

Financial Details

Financial Details:

Year	Payment type	Justification	Amount Requested
Year 1	Basic Salary	forecast for Lecturer 8.37 with a start date of 1/10/09 (12 months)	37,816.00
Year 1	On costs	NI + supperannuation	8,213.00
Year 1	Consumables	laptop+stationary	1,000.00
Year 1	Equipment		0.00
Year 1	Travel UK	1 conference (registration+accomodation+travel)	600.00
Year 1	Other Expenses		0.00
Total			47,629.00

Sum requested from the Royal Society: 47629.00

Start Date: 01/10/2009

Duration (Years): 1

Other expenses:

Peter Tiňo

30th January 2009

Senior Lecturer
School of Computer Science
The University of Birmingham
Birmingham B15 2TT, UK
(121) 414 8558
P.Tino@cs.bham.ac.uk
http://www.cs.bham.ac.uk/∼pxt/

RESEARCH AND TEACHING EXPERIENCE

- **January 2003 – : University of Birmingham, Birmingham, UK**
  Lecturer (Senior Lecturer since October 2006)
  Research: probabilistic modeling and machine learning for structured data, dynamical systems, evolutionary computation
  Teaching: Nature Inspired Learning, Neural Computation, Intelligent Data Analysis, Imaging and Visualisation Systems

- **March 2008 – September 2008**: **UK-Hong Kong Fellowship for Excellence**
  worked with Prof. H. Yan, City University of Hong Kong.
  Research: analysis of cDNA microarray data, promoter recognition.

- **May 2000 – December 2002: Neural Computing Research Group (NCRG), Aston University, Birmingham, UK**
  Postdoctoral research fellow, worked with Prof. I. Nabney
  Research: probabilistic modeling through latent variable models and data visualisation, machine learning in drug design
  Teaching: Practical Computation, Statistical Pattern Analysis

- **October 1997 – April 2000: Austrian Research Institute for Artificial Intelligence, Vienna, Austria**
  Postdoctoral research fellow, worked with Dr. G. Dorffner
  Research: multifractal analysis, connectionist models of natural language, modeling and analysis of financial data

- **October 1995 – September 1997: Slovak University of Technology**, Faculty of Electrical Engineering and Information Technology, **Bratislava, Slovakia**
  Assistant professor
  Research: recurrent neural networks, dynamical systems
  Teaching: Neural Networks, Modeling and Simulation, Computer Graphics

- **August 1994 – September 1995: NEC Research Institute, Princeton, USA**
  Fulbright fellow, worked with Prof. C.L. Giles
  Research: recurrent neural networks, grammar inference, dynamical systems

- **January 1992 – July 1994: Slovak University of Technology**, Faculty of Electrical Engineering and Information Technology, **Bratislava, Slovakia**
  Assistant professor

Research: recurrent neural networks, dynamical systems
Teaching: Neural Networks, Modeling and Simulation, Computer Graphics

- **October 1989 – December 1991: Slovak Academy of Sciences**, Institute of Control Theory and Robotics, **Bratislava, Slovakia**
  Study stay
  Research: fuzzy sets, belief functions, statistical inference

EDUCATION

- **May 1997: PhD**
  Institute of Control Theory and Robotics, Slovak Academy of Sciences, Slovakia
  Learning Temporally Dependent Associative Mappings with Recurrent Neural Networks

- **July 1988: MSc**
  Faculty of Electrical Engineering & Computer Science, Slovak University of Technology, Bratislava, Slovakia
  Thesis supervised by Dr. Zaťko, Comenius University, Faculty of Mathematics and Physics, Slovakia

HONOURS AND AWARDS

- **2008: UK/Hong Kong Fellowship for Excellence** (DfES)

- **2002: Best conference paper** award at **International Conference on Artificial Neural Networks (ICANN 2002)**
  P. Tiňo, B. Hammer: Architectural Bias in Recurrent Neural Networks - Fractal Analysis.

- **1998: Outstanding paper of the year** award for **IEEE Transactions on Neural Networks**
  T. Lin, B.G. Horne, P. Tiňo, C.L. Giles: Learning long-term dependencies with NARX recurrent neural networks. 1996

- **1994: Fulbright Fellowship**
  NEC Research Institute, Princeton, NJ, USA

- **1988: Graduation with Distinction**
  Slovak University of Technology, Slovakia

COOPERATION WITH INDUSTRY

- **2000 – 2003: Pfizer Central Research Ltd.**, UK
  applying machine learning techniques in drug discovery, searching for patterns in biological activity across multiple targets

PROGRAM COMMITTEE MEMBERSHIP

- 35 International conferences
- Technical Programme Committee Chair - 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07)

EDITORIAL BOARD

- Neural Processing Letters
- Pattern Analysis and Applications
- Advances in Artificial Intelligence
- The Open Artificial Intelligence Journal

JOURNAL REFEREEING

Extensive refereeing for many leading journals in the areas of machine learning, neurocomputing and physics including:
Neural Computation, Machine Learning, Neural Networks, Bioinformatics, Physica D: Nonlinear Phenomena, IEEE Transactions on Neural Networks, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Evolutionary Computation etc.

INVITED TALKS

22 invited talks including:
Carnegie Mellon University (2003), NIPS RNN workshop (2003), Oxford University (2004), British Computer Society Summer School on Pattern Recognition (annually since 2003), City University of Hong Kong (2008), Bielefeld University (2009), Dagstuhl seminars (2007, 2008, 2009).

Publications of Peter Tiňo

REFEREED PAPERS IN PRIMARY JOURNALS

1. H. Chen, P. Tiňo, X. Yao: Probabilistic Classification Vector Machines.
   IEEE Transactions on Neural Networks, accepted, 2009.

2. P. Tiňo: Bifurcation Structure of Equilibria of Iterated Softmax.
   Chaos, Solitons & Fractals, in print, 2009.

3. N. Gianniotis, P. Tiňo: Visualisation of tree-structured data through generative topographic mapping.
   IEEE Transactions on Neural Networks, 19(8), pp 1468-1493, 2008.

4. S.Y. Chong, P. Tiňo, X. Yao: Measuring Generalization Performance in Co-evolutionary Learning.
   IEEE Transactions on Evolutionary Computation, 12(4), pp 479-505, 2008.

5. P. Tiňo: Equilibria of Iterative Softmax and Critical Temperatures for Intermittent Search in Self-Organizing Neural Networks.
   Neural Computation, 19(4), pp. 1056-1081, 2007.

6. P. Tiňo, I. Farkas, J.van Mourik: Dynamics and Topographic Organization of Recursive Self-Organizing Maps.
   Neural Computation, 18(10), pp. 2529-2567, 2006.

7. J.C. Cuevas-Tello, P. Tiňo, S. Raychaudhury: How accurate are the time delay estimates in gravitational lensing?
   Astronomy and Astrophysics, 454(3), pp 695-706, 2006.

8. P. Tiňo, A. Mills: Learning Beyond Finite Memory in Recurrent Networks Of Spiking Neurons.
   Neural Computation, 18(3), pp. 561-613, 2006.

9. G. Brown, J. Wyatt, P. Tiňo: Managing Diversity in Regression Ensembles.
   Journal of Machine Learning Research, 6, pp. 1621-1650, 2005.

10. I. Nabney, Y. Sun, P. Tiňo, A. Kabán: Semisupervised Learning of Hierarchical Latent Trait Models for Data Visualization.
    IEEE Transactions on Knowledge and Data Engineering, 17(3), pp. 384-400, 2005.

11. P. Tiňo, I. Nabney, B.S. Williams, J. Losel, Y. Sun: Non-linear Prediction of Quantitative Structure-Activity Relationships.
    Journal of Chemical Information and Computer Sciences, 44(5), pp. 1647-1653, 2004.

12. G. Polčicová, P. Tiňo: Making sense of sparse rating data in collaborative filtering via topographic organization of user preference patterns.
    Neural Networks, 17(8-9), pp.1183-1199, 2004.

13. P. Tiňo, M. Čerňanský, L. Beňúšková: Markovian Architectural Bias of Recurrent Neural Networks.
IEEE Transactions on Neural Networks, 15(1), pp. 6-15, 2004.

14. P. Tiňo, B. Hammer: Architectural Bias in Recurrent Neural Networks - Fractal Analysis.
Neural Computation, 15(8), pp. 1931-1957, 2003.

15. B. Hammer, P. Tiňo: Recurrent neural networks with small weights implement definite memory machines.
Neural Computation, 15(8), pp. 1897-1926, 2003.

16. Ch. Schittenkopf, P. Tiňo, G. Dorffner: The Benefit of Information Reduction for Trading Strategies.
Applied Financial Economics, 34(7), pp. 917-930, 2002.

17. P. Tiňo: Multifractal properties of Hao's geometric representations of DNA sequences.
Physica A: Statistical Mechanics and its Applications, 304(3-4), pp. 480-494, 2002.

18. P. Tiňo, I. Nabney: Hierarchical GTM: constructing localized non-linear projection manifolds in a principled way.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), pp. 639-656, 2002.

19. P. Tiňo, Ch. Schittenkopf, G. Dorffner: Financial Volatility Trading using Recurrent Neural Networks.
IEEE Transactions on Neural Networks, 12(4), pp. 865-874, 2001.

20. P. Tiňo, Ch. Schittenkopf, G. Dorffner: Volatility Trading via Temporal Pattern Recognition in Quantized Financial Time Series.
Pattern Analysis and Applications, 4(4), pp. 283-299, 2001.

21. P. Tiňo, B.G. Horne, C.L. Giles: Attractive Periodic Sets in Discrete Time Recurrent Networks (with Emphasis on Fixed Point Stability and Bifurcations in Two–Neuron Networks).
Neural Computation, 13(6), pp. 1379-1414, 2001.

22. P. Tiňo, G. Dorffner: Predicting the future of discrete sequences from fractal representations of the past.
Machine Learning, 45(2), pp. 187-218, 2001.

23. P. Tiňo, M. Koteleš: Extracting finite state representations from recurrent neural networks trained on chaotic symbolic sequences.
IEEE Transactions on Neural Networks, 10(2), pp. 284-302, 1999.

24. P. Tiňo: Spatial Representation of Symbolic Sequences through Iterative Function Systems.
IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, 29(4), pp. 386-392, 1999.

25. T. Lin, B.G. Horne, P. Tiňo, C.L. Giles: Learning long-term dependencies with NARX recurrent neural networks.
IEEE Transactions on Neural Networks, 7(6), pp. 1329-1338, 1996.
Outstanding Paper of the Year award for IEEE Transactions on Neural Networks.

26. P. Tiňo, J. Šajda: Learning and Extracting Initial Mealy Machines with a Modular Neural Network Model.
Neural Computation, 7(4), pp. 822-844, 1995.

CONTRIBUTIONS TO SYMPOSIA AND COMPILED VOLUMES

27. X. Wang, P. Tino, M. Fardal: Multiple Manifold Learning Framework based on Hierarchical Mixture Density Model. In **Machine Learning and Knowledge Discovery in Databases (European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - ECML PKDD 2008)**, pp. 566-581, Lecture Notes in Computer Science, LNCS 4984, Springer-Verlag, 2008.

28. M. Čerňanský, P. Tiňo: Predictive Modeling with Echo State Networks . In **18th International Conference on Artificial Neural Networks - ICANN 2008**, Accepted. Lecture Notes in Computer Science, Springer-Verlag, 2008.

29. P. Tino: Bifurcations of Renormalization Dynamics in Self-organizing Neural Networks. In **14th International Conference on Neural Information Processing - ICONIP 2007**, (eds) M. Ishikawa et al. pp. 405-414, Lecture Notes in Computer Science, LNCS 4984, Springer-Verlag, 2008.

30. P. Tiňo, B. Hammer, M. Boden: Markovian bias of neural-based architectures with feedback connections.
In **Perspectives of Neural-Symbolic Integration**, (eds) B. Hammer, P. Hitzler. pp. 618-627, Studies in Computational Intelligence Vol. 77, Springer, 2007.

31. P. Tiňo, N. Gianniotis: Metric Properties of Structured Data Visualizations through Generative Probabilistic Modeling.
In **20th International Joint Conference on Artificial Intelligence - IJCAI'07**, (ed.) Manuela M. Veloso. pp. 1083-1088, AAAI Press, 2007.

32. P. Tiňo: On Conditions for Intermittent Search in Self-organizing Neural Networks. In **Advances in Artificial Intelligence - 6th Mexican International Conference on Artificial Intelligence - MICAI 2007**, (eds) A. Gelbukh, A. Fernando, K. Morales. pp. 172-181, Lecture Notes in Computer Science (4827), Springer-Verlag, 2007.

33. M. Čerňanský, P. Tiňo: Comparison of Echo State Networks with Simple Recurrent Networks and Variable-Length Markov Models on Symbolic Sequences.
In **17th International Conference on Artificial Neural Networks - ICANN 2007**, (eds) J. Marques de Sa, L.A. Alexandre, W. Duch, D.P. Mandic. pp. 618-627, Lecture Notes in Computer Science, Springer-Verlag, 2007.

34. N. Gianniotis, P. Tiňo: Visualisation of tree-structured data through generative probabilistic modelling.
In **15th European Symposium on Artificial Neural Networks - ESANN 2007**. pp. 97-102, 2007.

35. H. Chen, P. Tiňo, X. Yao: A Probabilistic Ensemble Pruning Algorithm.
In **6th IEEE International Conference on Data Mining - ICDM06 - Workshops**, (eds) . pp. 878-882, IEEE Computer Society, 2006.

36. J.C. Cuevas-Tello, P. Tiňo, S. Raychaudhury: A kernel-based approach to estimating phase shifts between irregularly sampled time series: an application to gravitational lenses.
In **17th European Conference on Machine Learning - ECML 2006**, (eds) J. Fuernkranz, T. Scheffer, M. Piliopoulou. pp. 614-621, Lecture Notes in Computer Science, Springer-Verlag, 2006.

37. P. Tiňo: Critical Temperatures for Intermittent Search in Self-Organizing Neural Networks.
In Parallel Problem Solving from Nature - PPSN IX, (eds) T.P. Runarsson, H-G Beyer, E. Burke, J J. Merelo-Guervos, L. Darrell Whitley, X. Yao. pp. 633-640, Lecture Notes in Computer Science, Springer-Verlag, 2006.

38. J.M. Binner, B. Jones, G. Kendal, J. Tepper, P. Tiňo: Does Money Matter? An Artificial Intelligence Approach.
In **Proceedings of 9th Joint Conference on Information Sciences 2006 (5th International Conference on Computational Intelligence in Economics and Finance)**, Kaohsiung, Taiwan. pp 72-75, 2006.

39. P. Tiňo, I. Farkaš, J.van Mourik: Recursive Self-Organizing Map as a Contractive Iterative Function System.
In **Intelligent Data Engineering and Automated Learning - IDEAL 2005**, (eds) M. Gallagher, J. Hogan, F. Maire. pp. 327-334, Lecture Notes in Computer Science, Springer-Verlag, 2005.

40. N. Nikolaev, P. Tiňo: Sequential Relevance Vector Machine Learning from Time Series.
In **Proc. Int. Joint Conference on Neural Networks - IJCNN 2005**, pp. 1308-1313, IEEE, 2005.

41. P. Tiňo, A. Mills: Learning Beyond Finite Memory in Recurrent Networks Of Spiking Neurons.
In **Advances in Natural Computation - ICNC 2005**, (eds) L. Wang, K. Chen, Y.S. Ong. pp. 666-675, Lecture Notes in Computer Science, Springer-Verlag, 2005.

42. P. Tiňo, I. Farkaš: On Non-Markovian Topographic Organization of Receptive Fields in Recursive Self-Organizing Map.
In **Advances in Natural Computation - ICNC 2005**, (eds) L. Wang, K. Chen, Y.S. Ong. pp. 676-685, Lecture Notes in Computer Science, Springer-Verlag, 2005.

43. P. Tiňo, N. Nikolaev, X. Yao: Volatility Forecasting with Sparse Bayesian Kernel Models.
In **Proceedings of 8th Joint Conference on Information Sciences 2005 (4th International Conference on Computational Intelligence in Economics and Finance)**, Salt Lake City, UT. pp 1150-1153 (on CD).

44. P. Tiňo, B. Hammer: On early stages of learning in connectionist models with feedback connections.
Compositional Connectionism in Cognitive Science – 2004 AAAI Fall Symposium Series, Washington, DC, 2004.

45. R. Price, P. Tiňo: Evaluation of Adaptive Nature Inspired Task Allocation Against Alternate Decentralised Multiagent Strategies.
In **Parallel Problem Solving from Nature - PPSN VIII**, (eds) X. Yao et al. pp. 982-990, Lecture Notes in Computer Science, Springer-Verlag, 2004.

46. G. Polčicová, P. Tiňo: Introducing a star topology into latent class models for collaborative filtering.
In **Proceedings of first IFIP Conference on Artificial Intelligence Applications and Innovations - WCC 2004**, (eds) M. Bramer, V. Devedzic. pp. 293-303, Kluwer academic publishers, 2004.

47. P. Tiňo, A. Kabán, Y. Sun: A generative probabilistic approach to visualizing sets of symbolic sequences.
The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (eds) R. Kohavi, J. Gehrke, W. DuMouchel, J. Ghosh. pp. 701-706, ACM Press, 2004.

48. Y. Sun, P. Tiňo, I. Nabney: Visualization of incomplete data using class information constraints.
In **Uncertainty in Geometric Computations**, (eds) J. Winkler, M. Niranjan. Kluwer, pp. 165-174, 2002.

49. P. Tiňo, Y. Sun, I. Nabney: Semi-Supervised Construction of General Visualization Hierarchies.
In **Proceedings of the 2002 International Conference on Artificial Intelligence - (IC-AI'02)**, (eds) H.R. Arabnia, Y. Mun. CSREA Press, pp. 1380-1386, 2002.

50. P. Tiňo, B. Hammer: Architectural Bias in Recurrent Neural Networks - Fractal Analysis.
In **Artificial Neural Networks - ICANN 2002**, (ed.) J.R.Dorronsoro. Lecture Notes in Computer Science, Springer-Verlag, pp. 1359-1364, 2002.
Best conference paper award.

51. A. Kabán, P. Tiňo, M. Girolami: A General Framework for a Principled Hierarchical Visualization of Multivariate Data.
In **International Conference on Intelligent Data Engineering and Automated Learning - IDEAL 2002**, Lecture Notes in Computer Science, Springer-Verlag, pp. 17-23, 2002.

52. P. Tiňo, M. Čerňanský, L. Beňúšková: Markovian Architectural Bias of Recurrent Neural Networks.
In **2nd Euro-International Symposium on Computational Intelligence**, (eds) P. Sincak, J. Vascak, V. Kvasnicka and J. Pospichal. IOS Press, pp. 17-23, 2002.

53. P. Tiňo, I. Nabney, Yi Sun, B.S. Williams: A Principled Approach to Interactive Hierarchical Non-Linear Visualization of High-Dimensional Data.
In **Computing Science and Statistics**, Volume 33, Proceedings of Interface '01 - Frontiers in Data Mining and Bioinformatics. 2002.

54. P. Tiňo, I. Nabney, Yi Sun: Using Directional Curvatures to Visualize Folding Patterns of the GTM Projection Manifolds.
In **Artificial Neural Networks - ICANN 2001**, (eds) G. Dorffner, H. Bischof and K. Hornik. Springer-Verlag, pp. 421-428, 2001.

55. P. Tiňo, G. Dorffner, Ch. Schittenkopf: Understanding State Space Organization in Recurrent Neural Networks with Iterative Function Systems Dynamics.
In **Hybrid Neural Symbolic Integration**, (eds) S. Wermter, R. Sun. pp. 256-270, Springer Verlag, 2000.

56. P. Tiňo, Ch. Schittenkopf, G. Dorffner: Methods of Symbolic Dynamics in Options Trading.
In Proceedings of **Computational Finance 2000**, London, UK (on CD).

57. P. Tiňo, M. Stančik, L. Beňušková: Building predictive models on complex symbolic sequences with a second-order recurrent BCM network with lateral inhibition.
In Proceedings of the **IEEE-INNS-ENNS Int. Joint Conference on Neural Networks**, Como, Italy. Vol. 2, pp. 265-270, 2000.

58. P. Tiňo, M. Stančik, L. Beňušková: Building predictive models on complex symbolic sequences via a first-order recurrent BCM network with lateral inhibition.
In **Quo Vadis Computational Intelligence? New Trends and Approaches in Computational Intelligence**, (eds) P. Sinčák and J. Vaščak. Physica-Verlag, Heidelberg, pp. 42-50, 2000.

59. Ch. Schittenkopf, P. Tiňo, G. Dorffner: The profitability of trading volatility using real-valued and symbolic models.
Proceedings of the **IEEE/IAFE conference on Computational Intelligence in Financial Engineering - CIFEr 2000**, New York City, NY, USA. pp. 8-11, 2000.

60. Sh. Parfitt, P. Tiňo, G. Dorffner: Graded grammaticality in Prediction Fractal Machines.
In **Advances in Neural Information Processing Systems - NIPS 12**, (eds) S. A. Solla, T. K. Leen, K-R. Mller. pp. 52-58, MIT Press, 2000.

61. P. Tiňo, G. Dorffner: Building predictive models from spatial representations of symbolic sequences.

In **Advances in Neural Information Processing Systems - NIPS 12**, (eds) S. A. Solla, T. K. Leen, K-R. Mller. pp. 645-651, MIT Press, 2000.

62. P. Tiňo, Ch. Schittenkopf, G. Dorffner, E.J. Dockner: A Symbolic Dynamics Approach to Volatility Prediction.
In **Computational Finance**, (eds) Y.S. Abu-Mostafa, B. LeBaron, A.W. Lo, A.S. Weigend. pp. 137-151, MIT Press, Cambridge, MA, 2000.

63. T. Lin, B.G. Horne, P. Tiňo, C.L. Giles: Learning Long-Term Dependencies in NARX Recurrent Neural Networks.
In **Recurrent Neural Networks - Design and Applications**, (eds) L.R. Medsker, L.C. Jain. pp. 133-152, CRC Press, 1999.

64. P. Tiňo, V. Vojtek: Modeling complex sequences with recurrent neural networks.
In **Artificial Neural Networks and Genetic Algorithms**, (eds) G.D. Smith, N.C. Steele, R.F. Albrecht. pp. 459-463, Springer Verlag, 1998.

65. P. Tiňo, B.G. Horne, C.L. Giles, P.C. Collingwood: Finite State Machines and Recurrent Neural Networks – Automata and Dynamical Systems Approaches.
In **Neural Networks and Pattern Recognition**, (eds) J.E. Dayhoff, O. Omidvar. pp. 171-220, Academic Press, 1998.

66. P. Tiňo, V. Vojtek: Spatial Representation of Temporal Structure in Symbolic Sequences through Iterated Function Systems.
In Proceedings of the **International Conference on Measurement - MEASUREMENT'97**, (eds) Ivan Frollo, Anna Plackova. 1997.

67. T. Lin, B.G. Horne, P. Tiňo, C.L. Giles: Learning long-term dependencies is not as difficult with NARX recurrent networks.
In **Advances in Neural Information Processing Systems - NIPS 8**, (eds) D.S. Touretzky, M.C. Mozer, M.E. Hasselmo. pp. 577-602, MIT Press, 1996.

68. P. Tiňo, M. Koteleš: Modeling Complex Symbolic Sequences with Recurrent Neural Networks.
In Proceedings of the **1-st Slovak Neural Network Symposium**, pp. 78-85, 1996.

69. P. Tiňo, B.G. Horne, C.L. Giles: Stability and bifurcations analysis of fixed points in discrete time recurrent neural networks with two neurons.
In Proceedings of the **World Congress on Neural Networks**, Washington D.C., Vol 3, pp. 170-173, 1995.

70. T. Lin, B.G. Horne, P. Tiňo, C.L. Giles: Long-term dependencies in NARX networks.
In Proceedings of the **World Congress on Neural Networks**, Washington D.C., Vol 3, pp. 142-146, 1995.

71. P. Tiňo, I.E. Jelly, V. Vojtek: Non-Standard Topologies of Neuron Field in Self-Organizing Feature Maps.
In Proceedings of the **AIICSR'94** conference, pp. 391-396, World Scientific Publishing Company, 1994.

BOOKS

72. H. Yin, P Tiňo, E. Corchado, W. Byrne, X. Yao (Eds.): **Intelligent Data Engineering and Automated Learning - IDEAL 2007**. Springer, Lecture Notes in Computer Science, Vol. 4881, 2007.

73. X. Yao, E. Burke, J.A. Lozano, J. Smith, J.J. Merelo-Guervs, J.A. Bullinaria, J. Rowe, P. Tiňo, A. Kabán, H.P. Schwefel (Eds.): **Parallel Problem Solving from Nature - PPSN VIII**.
Springer, Lecture Notes in Computer Science, Vol. 3242, 2004.

74. V. Kvasnička, J. Pospichal, P. Tiňo: **Evolutionary algorithms**. (in Slovak).
STU, Bratislava, 2000.

75. V. Kvasnička, L. Beňušková, J. Pospichal, I. Farkaš, P. Tiňo, A. Král: **Introduction to the Theory of Neural Networks**. (in Slovak).
IRIS, Bratislava, 1997.

A New Framework for Holistic Representations and Learning on Aminoacid Sequences

Peter Tiňo

Background

We are witnessing a revolution in biology driven by the large-scale data amassed by current experimental technologies and efforts to uncover the function of whole-genome data. Proteomics data (reflecting the state of a cell's protein content) and next-generation sequencing data will very likely revolutionise our insights into genomics at the level of individuals, with a huge potential for personalised health care. However, for these exciting developments to take place, we need dedicated tools enabling exploration of such large-scale data. In this proposal we will concentrate on situations were the data represent complex entities characterised by an underlying sequential structure endowed with a variety of structural, physical and chemical properties that determine their biological function. All these factors should be holistically taken into account when learning on such data. For example, proteins from one point of view are just long molecules of chained aminoacids. But proteins can also form complicated localised (secondary structure) and global (tertiary structure) spatial arrangements. In addition, each aminoacid (or small group of aminoacids) has special physical-chemical properties that together with the spatial arrangements of the protein are key factors determining the protein function.

Previous attempts to extend the sequential information with additional properties characterising proteins have been dominated by kernel machines and data representations through long feature vectors, typically encompassing only a limited number of feature types (e.g. (important sequential motifs, localised physical properties)). We propose to explore a radically different view on data analysis and automated learning in the context of biological aminoacid sequences. We allow to capture a potentially large variety of protein features in a natural algebraic structure of higher-order tensors. We then perform a *task-driven compression* of the data tensors *in the tensor space*, thus achieving a lower-rank representation of the dominant trends in the data set that are crucial for the given task. If we used traditional vector representations, such a compression would be problematic due to the vast number of degrees of freedom involved. As the nature of the task changes, so does the view on what are the important and dominant factors in the structure-rich data. Making the learning machines operate on the compressed space may not only boost their generalisation performance but crucially can also increase their interpretability.

We will verify the framework on a representative problem of nuclear protein localisation. Understanding nuclear localisation of proteins (e.g. the process by which proteins get transported into the cell nucleus) is a crucial step in understanding the dynamics and self-regulation of the cell. Briefly, the transport of most molecules between the cytoplasm and nucleus through the so called Nuclear Pore Complex is assisted by specialised importins and exportins. Importins recognise molecules to be imported into the nucleus through Nuclear Localisation Signals (NLS). However, the localisation signals are extremely varied. Some localisation signals are aminoacid subsequences that can potentially appear anywhere in the sequence, but are physically exposed on the surface of the 3-dimensional folded protein structure. Many other localisation signals have a much more complicated structure. In addition, NLS by themselves are not sufficient to resolve the localisation issue as there are many known non-nuclear proteins containing NLS and there are known nuclear proteins that do not contain any NLS. The situation is further complicated by the existence of other nuclear localisation mechanisms, the known range of which is constantly increasing. Last but not least, some nuclear proteins can be dually (or even multiply) localised, for example proteins shared between the nucleus and cytoplasm in a shuttling process. To date there are only three specialised predictors for identifying nuclear proteins and several more general predictors of protein localisation many of which rely on sequence homology (e.g. common evolutionary history of the studied proteins). However, models exploiting sequence homology are unable to deal with novel protein sequences far removed from the known ones. Furthermore, except for NUCLEO predictor, dually localised proteins have so far not been considered at all. To summarise, predicting nuclear localisation is an important but very difficult task. We expect that using compressed features of the detailed tensor-based data representations will help to reveal new findings about the nature and mechanisms of nuclear protein localisation that have not been accessible through the data representation schemes used so far.

Aims, methodology, timeline and milestones

There are two aims (coinciding with the tasks) of the proposed work. The milestones correspond to the end of each task.

1. Develop a framework for learning models based on tensor representation of various modes of information characterising proteins. (8 months)

2. Based on such framework investigate in the supervised and unsupervised learning setting the signalling mechanisms of singular as well as dual nuclear protein localisation. (4 months)

Methodology - Task 1: If it is possible to define a positional reference frame for the set of proteins to be explored, the first information mode will be the position within the aminoacid sequence. Important phrases/motifs will be represented along mode 2, local secondary structure along mode 3 and physical-chemical properties of the relevant aminoacids along mode 4. For example, when considering W distinguishing words, S secondary structure types, P property types, then length-L cuts of proteins will be represented as fourth-order $L \times W \times S \times P$ tensors

with element (i, w, s, p) equal to 1 if at position i within the sequence, there is the word w, local structure s and property p. Otherwise the (i, w, s, p)-th element of the representational tensor will be 0. If a positional reference frame cannot be defined, the mode 1 will be dropped, resulting in order-3 tensors having the (w, s, p)-th element equal to 1 if within the protein there is the word w at a position corresponding to the local structure s and property p. Additional modes can be added analogously.

Such protein representation will code richer information than most of the current approaches to learning on aminoacid sequences. To account for the binary nature of data tensors, each tensor element will be modelled by a Bernoulli noise distribution. However, the representations will be high-dimensional and sparse. To extract the dominant trends in such data, compression in the tensor space will be performed. Several criteria will be used to drive the compression. We will impose specific constraints on the noise models so that the number of degrees of freedom is dramatically reduced, while retaining a good model of the data. The constraining mechanisms include: **(i)** Imposing that the natural parameters of the individual Bernoulli models lie in a low-dimensional linear tensor subspace; **(ii)** the low-dimensional tensor basis can be factored into separate basis vectors along each mode; **(iii)** allowing for possible controlled non-linear structure of the tensor subspace by mapping the subspaces of (i-ii) via a smooth parametrised non-linear mapping. The task that needs to be performed on the proteins will drive finding the reduced tensor basis. For example, if the proteins are to be classified into separate classes (e.g. nuclearly localised vs. non-localised proteins), the model will find the tensor basis that can best separate the data into the classes (supervised learning). Analogously, in the unsupervised learning setting, if the task is to find natural groupings of the proteins, the tensor basis will best separate and compactify the data clusters. If the goal is to visually represent the proteins in a topographic map, we would aim for the smallest reconstruction error under the reduced basis. Furthermore, we will account for tensor sparsity by building in the option for allowing a higher model construction cost for miss-modelling 1 than for miss-modelling 0.

Methodology - Task 2: The data will be prepared using the latest release of Swiss-Prot and we will apply stringent redundancy reduction to drive the models to genuine generalisation not relying on inherent sequence homologies. Proteins not localised to nucleus will be represented by those with known subcellular localisation to an organelle other than nucleus. The set of known localisation motifs will be extended by the set of most distinguishing words between the localised and non-localised protein classes. The latter will be found using a suffix tree representation. The methodology enables finding variable length word patterns (as opposed to fixed length patterns used in most representational schemes). Starting from short patterns, each pattern is extended only if the extended version has a more distinguishing power, subject to blocking inclusion of low-frequency patterns to avoid overspecialisation.

We will systematically examine combinations of mode types and mode numbers starting with tensors of order 2, moving to the order-3 tensors etc. The unsupervised and supervised learning investigations will differ in the data representation process (e.g. finding specific vocabularies of the most distinguishing words), the reduced tensor basis construction and the learning machinery on the reduced protein representations. In the unsupervised learning setting, we will use clustering and topographic mapping to detect natural groupings of nuclear proteins revealed by commonalities in the compressed tensor representations and translate them into a biological interpretation. In the supervised mode, the models will be used to classify proteins into nuclear localised vs. non-localised, or single localised vs. dually localised proteins. Where possible, the prediction performance will be stringently compared with the state of art alternatives. Successful models will be analysed to unveil the features that most separate the protein classes and the features will be verified for their biological meaning. The biological interpretation of the results will be done in close collaboration with Dr. Mikael Bodén and his colleagues at the Institute of Molecular Bioscience, University of Queensland.

Why the Theo Murphy Blue Skies awards Scheme?

This truly interdisciplinary research will bridge machine learning, bioinformatics and biology. The key idea of this project is to allow for sparse tensor representations of protein features, potentially far richer than the ones currently in use in learning machines operating on aminoacid sequences. Our intuition is that the sparsity can be dealt with by appropriate compression in the tensor space, respecting the sparse and binary nature of the tensors. Moreover, the dominant trends in the task-driven compressed representations will constitute hypothesis that will be verified for biological significance. While such an approach can bring significant advances in automated learning on protein data, it may also be the case that the tensor representations will not provide enough structure for an effective compression to take place. In that case we will systematically search for the minimal mode subset leading to lower-order tensors on which automated learning can be successfully performed. If that fails, to prove usefulness of incorporating a wide variety of features capturing the proteins, we will resort to fully kernel machines based approaches and construct dedicated protein kernels operating on aminoacid sequences, but extended with as much additional information as possible. If successful, our proposal can make a high impact in the bioinformatics field, but needs a proof of concept before further major developments can take place. Relying on task-driven compression of vast sparse binary tensors representing proteins in a holistic manner is a novel concept without sufficient evidence base and as such would be considered as lacking feasibility component by traditional grant schemes.

Relevant research experience

I have an extensive research experience in learning machines operating on structured data (predominantly sequential) and probabilistic modelling (especially topographic mapping and dimensionality reduction via latent variable modelling). In both fields I have published in top scientific journals. Directly related to this proposal is my work on topographic maps and clustering of high-dimensional and structured data in the following publications (numbering taken from my publication list submitted separately) [3,6,10,12,18,31,46-49,51,54]. I have maintained a keen interest in **interdisciplinary research**: **Drug discovery** (Pfizer Research) - machine learning in drug discovery (2000 2003); **Astronomy** (School of Physics and Astronomy, University of Birmingham) - machine learning for resolving time delays in gravitational lensing, clustering and topographic mapping of binary star complexes (2003-); **Bioinformatics** (City University of Hong Kong) - analysis of cDNA microarray data, promoter recognition (2008)