Advances in the Application of Machine Learning Techniques in Drug Discovery, Design and Development

S.J. Barrett† and W.B. Langdon*

† Analysis Applications, Research and Technologies, GlaxoSmithKline
R&D, Greenford Rd, Greenford, Middlesex, UB6 0HE. UK
*Computer Science, University of Essex, Colchester CO4 3SQ, UK

Abstract. Machine learning tools, in particular support vector machines (SVM), Particle Swarm Optimisation (PSO) and Genetic Programming (GP), are increasingly used in pharmaceuticals research and development. They are inherently suitable for use with 'noisy', high dimensional (many variables) data, as is commonly used in cheminformatic (i.e. In silico screening), bioinformatic (i.e. bio-marker studies, using DNA chip data) and other types of drug research studies. These aspects are demonstrated via review of their current usage and future prospects in context with drug discovery activities.

1 Introduction

Pharmaceutical discovery and development is an evolving [Ratti & Trist, 2001] cascade of extremely complex and costly research, comprising many facets [Ng, 2004] which create a vast diversity of data and sub-problems [Butte, 2002; Schrattenholtz, 2004; Watkins & German, 2002; Roses, 2002]. Drug design and optimisation increasingly uses computers [Hou and Xu, 2004; Schneider and Fechner, 2005] and more commonly against vast 'integrated' research datasets constructed from large inhomogeneous combinations of data (from disparate sources and disciplines) to answer novel lines of inquiry, and for the generation of research hypotheses.

Conventional statistical methods are currently better known and understood by Pharmaceuticals R&D scientists who benefit from the traditional statistical support toward design of experiments, data assessment, etc. However statistical groups are increasingly using other computational methods and recognising alternative approaches [Hand, 1999; Breiman, 2001], as existing (usually hypothesis testing) methods are found lacking. This is generally due to the increasing need for data exploration and hypothesis generation in the face of growing data, problem complexities, and *ad hoc* experimental design inadequacies from compromises due to cost and lack of prior knowledge. As individual techniques may only partly cope with these problems, multiple methods are often used for comparative analyses. Whilst conventional multivariate statistical methods remain of great utility, most are inherently linear lessening their suitability for a plethora of newer, more complex problems. Consequently, evaluation and early uptake of novel computational approaches continues within pharmaceuticals research, with scientists increasingly turning to recursive partitioning, Artificial Neural Networks (ANNs) and other methods. Whilst ANNs and genetic algorithms are established in traditional application areas [Jones, 1999; Solmajer and Zupan, 2004], tracking the uptake of more recent machine learning approaches is difficult due to the diversity of new applications and fragmented literature.

The newer predictive modeling approaches include Support Vector Machines (SVM) and evolutionary computing paradigms such as Genetic programming and Particle Swarm Optimisation (PSO). SVM algorithms arose [Boser *et al.*, 1992] from concepts of structural risk minimisation and statistical learning theory [Vapnik, 1995]. SVMs are commonly used for classification (SVC) and regression (SVR). They are a sophisticated synthesis of ANN-like hyperplane methodology, backed by a sound theory of learning and convergence, applying robust linear methods (and within kernel spaces for non-linear classifiers) to give excellent generalisation characteristics [Shawe-Taylor and Cristianini, 2000]. In contrast to the rigorous mathematical approach of SVMs, genetic programming [Banzhaf, *et al.*, 1998] appeals to metaphor. GP uses Darwins' natural selection to evolve a population of computer programs. The better programs are selected to be parents for the next generation. Children are created by crossover and mutation. Some are better and some are worse than their parents. Selection continually encourages better individuals to pass on their genes. Overtime and successive generations the population improves until an individual with satisfactory performance is found. Many drug discovery problems can be expressed as the problem of finding a computer program, and GP is general purpose requiring minimal assumptions and capable of solving very difficult problems. Particle Swarm Optimization (PSO) was inspired by swarms of insects, shoals of fish, etc [Eberhart, Kennedy and Shi, 2001]. In PSO's the creatures are abstracted to moving particles. These fly over the problem space. If they find a good point they are randomly attracted back to it. Fundamentally PSO also have a similar social force which attracts all the particles to the best solution to the problem found by the whole swarm.

We here review the current status of pharmaceutically relevant applications of Support Vector Machines (SVMs), Genetic Programming and Particle Swarm Optimisation and assess briefly assess their future.

2 SVM Applications in Pharmaceuticals Research

2.1 SVM in Cheminformatics and Quantitative Structure-Activity Relationship (QSAR) Modelling.

Cheminformatics in drug discovery has been reviewed by [Xu and Hagler, 2002]. An early task is the creation of virtual respresentations of molecules and assessment of their likely suitability for synthesis and viability for development for use in the body. SVC predictions of 'drug-likeness' from virtually-represented molecules are reportedly more robust than those from ANNs [Byvatov et al. 2004], achieving success in predicting chemists' intuitive assessments [Takaoka et al., 2003]. Cheminformatics combines chemical properties and high-throughput screening measurements, in large scale QSAR. Trained SVM-QSAR classifiers now enable 'virtual screening' for discovering molecules with specific therapeutic target affinities from millions of virtual representations [Jorissen and Gilson, 2005], reducing the scale of subsequent 'physical' screening of synthesised molecules. SVC 'active learning' has been used to reduce the number of drug-optimising synthesis-biotesting cycles [Warmuth et al., 2003]. Studying bio-active conformations of molecules aids understanding of mechanisms of action for improving specificity and selectivity and [Byvatov et al., 2005b] have used SVM methodology to study molecular pharmacophore patterns. [Chen, 2004] reports on SVM uses in the wider field of chemistry.

Predicting activity toward specific therapeutic targets. G-protein coupled receptors (GPCRs) are the major class of drug targets. [Suwa *et al.*, 2004] provided physicochemical features of GPCRs and their ligands to a Radial Basis Function-SVC (RBF-SVC) to predict specific G-protein couplings. [Cheng *et al.*, 2004] used an RBF-SVR to predict antagonist compound metabolism and inhibitory activity toward human glucagon receptor to select 3D QSAR features. [Byvatov, *et al.*, 2005] used binary SVC active learning to enrich dopamine receptor agonists, applying SVR to the enriched set to predict D2/ D3 receptor selectivity. [Takahashi *et al.*, 2005] used multiclass SVC to predict D1 receptor agonists, antagonists and inactives. [Burbidge, 2004] applied SVM to a variety of monoamine QSAR problems, but good performance could come with non-sparsity: a large number of training points as support vectors can severely reduce prediction speed in virtual screening. [Burbidge *et al.*, 2001a] devised an algorithm to counter this.

Predicting Absorption Distribution Metabolism Excretion Toxic (ADMET) effects. [Burbidge, *et al.*, 2001b] favourably compared SVC to ANNs, decision trees and K-nearest-neighbour (k-NN) classifiers for predicting human blood-brain barrier penetration, human oral bioavailability and protein-binding. [Brenemann *et al.* 2003] applied SVM to cell permeability prediction. Bacterial P-glycoprotein (P-gp) mediated efflux of substrate antibiotics results in drug resistance. [Xue *et al.*, 2004a] used Gaussian SVC Recursive Feature Elimination (SVC-RFE) to predict P-gp substrates, outperforming ANN and k-NN. [Xue *et al.*, 2004b] used similar approach for predicting human intestinal absorption and serum albumin binding. [Doniger *et al.*, 2002] demonstrated benefits of RBF-SVC over ANNs against a small dataset to predict central nervous system (Blood-Brain Barrier) permeability. [Norinder, 2003] had overfitting problems with SVR, needing simplex optimization for parameter and feature selection to achieve good predictors for BBB penetration and human intestinal absortion. [Liu *et al.*, 2005] used Gaussian SVR for predicting human oral drug absorption.

[Yap *et al.*, 2004] used Gaussian SVC to differentiate drugs that can cause *torsade de pointes* (TdP), an adverse drug reaction which involves multiple mechanisms. Prediction accuracy compared favourably with k-NN, ANN and C4.5. [Xue *et al.*, 2004b] also used SVC, but with RFE to predict TdP inhibition. [Tobita, *et al.*, 2005] used RBF-SVC to predict chemical inhibition of *HERG* potassium channel that is associated with heart arrhymia which can trigger TdP. Non-Steroidal Anti-Inflammatory Drugs reduce inflammation by blocking cyclo-oxygenase enzymes and selective blocking of the COX-2 form reduces gastro-intestinal side effects. [Liu, *et al.*, 2004] employed RBF SVC/SVR to discriminate between COX inhibitors.

Cytochrome p450 (CYP) enzymes are important chemical (and drug substrate) metabolisers within the body, and significant drug inhibition of these is to be avoided. Superior prediction of CYP3A4 inhibition has been reported with SVC compared to other methods [Merkwirth *et al.*, 2004; Arimoto & Gifford, 2005]. SVM methods have also been used to predict CYP2D6, CYP2C9 [Yap & Chen, 2005] and CYP1A2 inhibition [Kless & Eitrich, 2004].

2.2 SVM in Bioinformatics

SVM application in bioinformatics has been reviewed by [Byvatov and Schneider, 2003]. Here we present an update.

Gene Expression Micro-Array Data in the Prediction of Disease Traits. As with SNPs data, input dimensionality can be extremely large (10Ks of genes) whilst the number of examples is relatively small (typically 10s to 100s). Whilst SVMs are relatively well suited to this situation, [Malossini *et al.*, 2004] showed significant performance degradation with just a few incorrectly labelled training examples (as can occur in complex disease diagnosis). Large numbers of correlated and irrelevant genes also diminish performance, making feature selection essential. [Guyon *et al.*, 2002] invented Recursive Feature Elimination (RFE), employing SVC within a wrapper-based approach although [Ambroise and Mclachan, 2002] reported gene selection bias with this. Related 'entropic' [Furlanello *et al.*, 2003] and Recursive Feature Replacement (RFR) [Fujarewicz and Wiench, 2003] followed outperforming earlier methods, with RFR best for smaller gene subsets [Simek *et al.*, 2004].[Fung and Mangasarian, 2004] have achieved sparse models directly with fast linear programming SVC. SVCs are regularly used to predict cancer cases using gene expression training data [Wang *et al.*, 2005], and chemo-genomic studies (of functional relationships between genes and drugs) are also increasing [Bao and Sun, 2002; Thukral *et al.*, 2005].

Receptor Classification and Protein Function Annotation. SVM methods are now often employed to predict the functional classes of proteins from sequence data, i.e. GPCR families or nuclear receptor sub-family [Bhasin and Raghava, 2004a,b] and enzyme class [Dobson and Doig, 2005].

Gene Functional Classes and Annotation. [Brown *et al.*, 2000] first employed SVC to predict functional classes of genes, others have continued in this vein, i.e. [Vinayagam *et al.*, 2004] devised a large-scale gene annotation system exploiting the gene-ontology DAG structure using multiple SVCs for prediction correctness.

Proteomics/Protein Expression. [Jong *et al.*, 2004] Studied predictability of prostate and ovarian cancers using SELDI-TOF mass spectronomy (MS), achieving excellent performance with linear SVC. [Seike *et al.*, 2004] used SVC within a methodology to rank protein spots (in expression profiles from 2D-gel electrophoresis) in terms of their discrimatory ability for human cancers. [Prados *et al.*, 2004] found linear-SVC to out-perform k-NN, ANN and decision tree approaches in predicting ischemic and haemorrhagic stroke from SELDI-MS data applying weight interrogation to identify candidate biomarkers. [Bock and Gough, 2003] used SVC in a system generating protein-protein interaction hypotheses for constructing protein interaction networks.

Other Bioinformatics Applications. [Schneider and Fechner, 2004] have reviewed machine learning approaches (including SVMs) to protein sub-cellular localisation for target identification in drug discovery. There is a growing use of SVC prediction of functionally critical sites within proteins, i.e. sites of: phosphorylation [Kim *et al.*, 2004], ATP-binding [Guo, *et al.*, 2005], catalysis [Dubey *et al.*, 2005] and cleaving [Yang and Chou, 2004]. Specialist kernels have arisen here, i.e. for protein homology [Saigo *et al.*, 2004]) and siRNA design for 'gene-silencing' [Teramoto *et al.*, 2005].

2.3 SVM in Clinical Diagnosis and Epidemiology

Molecular Genetic Epidemiology. Single-Nucleotide Polymorphisms (SNPs) are common individual base changes within human DNA. Millions have been identified. Unlike gene expression measures, SNPs represent unchanging patient-specific variation that may relate to an individuals' prognosis. The feasibility of using SVC methodology to predict disease using multiple SNP variations has been demonstrated for coronary heart disease [Yoon *et al.*, 2003] and breast cancer [Listgarten *et al.*, 2004]. [Barrett, 2005] used SVC to find SNPs associated with drug effect via iterative training and SNP-removal using 1-norm linear SVC weight-vector interrogation.

Epidemiology and Clinical Diagnostics. Apart from in the 'molecular-related' contexts (as above) the use of SVM in epidemiology remains in its infancy. Observing that variable interactions are often not considered in standard univariate analyses, [Fradkin, 2005] discusses the potential of SVM models to provide an alternative to the standard logistic regression method used to identify risk factors in cross-sectional studies. In the only reported study of SVM modelling of large epidemiological observational data, [Muchnik, 2001] used the SEER database, computing multiple SVC models (using variable perturbation) to identify candidate epidemiological factors influencing on breast cancer survival time.[Härdle and Moro, 2004] used SVM to achieve breast cancer survival analysis. [Zhao *et al.*, 2004] used SVC to differentiate anorexic patients. There is a much wider use of SVC in clinical diagnostics with large complex data from sophisticated equipment such as EEG (epilepsy: [Miwakeichi *et al.*, 2001]; CT (colon cancer: [Jerebko, *et al.*, 2005]), MRI (brain glioma: [Li *et al.*, 2005]) and sonography (breast cancer: [Huang & Chen, 2005]).

3 Drug Research Applications of Genetic Programming

In most Pharmaceutical applications, GP evolves predictive models. Typically these take data (i.e. number of positively charged ions, presence of aromatic rings, , etc.) and predict whether a molecule inhibits an enzyme or not. There are now at least two annual workshops on EC uses in Biology: BioGEC (2002-06) and EvoBIO (2003-06).

3.1 GP in Cheminformatics and QSAR.

GP has been used for combinatorial design [Nicolotti et al., 2002], modelling drug bioavailabity [Langdon et al., 2002] and HERG inhibition [Bains et al., 2004], whilst ensembles of ANNs have been evolved to predict p450 inhibition [Langdon et al., 2002a].

3.2 GP in Bioinformatics.

Hot topics include: DNA and protein sequence alignment [Shyu et al., 2004]; protein localisation [Heddad et al., 2004]; using genetic algorithms etc. to infer phylogenetics trees [Congdon and Septor, 2003]; classification and prediction [Hong and Cho, 2004]; recognising transmembrane regions of proteins [Koza and Andre, 1996]; and finding DNA promoters [Howard and Benson, 2003] and gene regulatory sites. Infrared spectroscopy, DNA chip and Single Nucleotide Polymorphisms (SNPs) [Reif et al., 2004] datasets have huge numbers of features. Often the immediate problem is to discover which of the thousands are relevant. In [Johnson et al., 2003] isolation of the relevant wave numbers using GP revealed new insights into commercial crops. GP has also been used to sift thousands of inputs in DNA chip data to discover which genes are important to a metabolic process [Langdon and Buxton, 2004; Moore et al., 2002] or to reduce the number of inputs required so a diagnostic test is practicable [Deutsch, 2003]. While GAs can achieve high multi-class accuracy [Ooi and Tan, 2003] they are also commonly combined with other classifiers, e.g. linear [Smits et al., 2005], SVM [Li et al., 2005], naive Bayes [Ando and Iba, 2004] and k-nearest neighbour. It is no wonder that GP is increasingly being used in Bioinformatics data mining [Kell, 2002]; modelling genetic interactions [Moore and Hahn, 2004] and organisms; inferring metabolic pathways [Koza et al., 2001; Tsai and Wang, 2005] and gene regulatory networks.

3.3 GP in Clinical Diagnosis and Epidemiology Research.

So far, GP is not so used, although GP has been applied to diagnosing pulmonary embolism [Biesheuvel, 2005] and atherosclerosis risk [Sebag et al., 2004].

4 Biological Applications of Particle Swarm Optimisation

Unlike GP, the current use of PSOs in pharmaceutical research is relatively unexplored. Commonly PSOs are used in hybrids with other approaches. PSOs naturally search widely, making them suited to finding good regions. Exploitive local method is then used to refine the good starting points found by PSOs into excellent solutions.

4.1 PSO in Cheminformatics and QSAR.

In QSAR a few teams have used a two stage approach. In the first stage a binary PSO is used to select a few (typically 3-7) features as inputs to supervised learning method. In [Lu *et al.,* 2004] the BPSO selects 7 of 85 features. Then linear models of drug activity (IC50) with two enzymes, COX-1 and COX-2, are constructed. (In [Lin *et al.,* 2005] they use a PSO to divide low dimensional, e.g. 5 features, chemical spaces into pieces. A linear model is fitted to each sub-region.). To aid *in silico* design of drugs, [Lu *et al.,* 2004] produce models which may differentiate between binding to the two enzymes by (virtual) chemicals.

[Wang *et al.,* 2004] and [Shen *et al.,* 2004] use feed-forward ANN to classify the Bio-activity of chemicals using a few (3-6) features selected by a BPSO. They also consider replacing the ANN by a k-nearest neighbour classifier in combination with kernel regression. While they note some differences, many approaches turn out to be equally good at predicting which chemicals will be carcinogenic. The datasets typically only cover a few (31-256) chemicals but, for each one, a large number (27-428) of features are computed from its chemical formula. One can reasonably argue that some form of "feature selection", i.e. choosing which attributes can be used by the ANN, is essential. Even so, given the small number of chemicals involved, [Agrafiotis and Cedeno, 2002; Cedeno and Agrafiotis, 2003; Wang *et al.*, 2004] are still careful to prevent over fitting, e.g. by the use of "leave-n-out" cross-validation.

4.2 PSO in Bioinformatics.

DNA chip experiments often mean under-constrained biomarker search problems (many variables vs few examples). [Xiao et al., 2003] use self organising maps (SOM) to pick clusters of similar genes from datasets with thousands. The PSO is seeded with crude SOM results to refine the clusters.

4.3 PSO in Clinical Diagnosis and Epidemiology Research.

Two and three dimensional medical images, such as X-Rays and MRI, can contain millions of data per subject. [Wachowiak et al., 2004] propose a hybrid PSO to match images taken at different times and/or with different techniques (e.g. ultrasound, CT). Best results came by combining expert medical knowledge to give an initial alignment and a PSO. [Eberhart and Hu, 1999] used a PSO to train an ANN which, using wrist accelerometer data, identifies essential tremor and Parkinson's disease sufferers.

5 Discussion

Whilst the above survey clearly demonstrates a wide coverage of relevant problem areas, it remains unclear as to the underlying extent to which these approaches are actually deployed across the pharmaceuticals industry so their overall importance there is difficult to ascertain. Although becoming less sporadic, it seems that the use of machine learning is still largely driven by individuals either with their own expertise and/or external expert resources.

Machine learning has however proved its worth in many areas for fundamental reasons (for instance model transparency is a recognized benefit of evolutionary methods and SVMs are well known for their generalization). For these newer technologies to make further applications advances there is a need for ease-of-use; easier derivation of problem-specific representations; adequate ways of handling missing data; more widespread generation of reliable prediction confidence measures and attention to statistical power of datasets in model selection. Encouragingly, the machine learning research community is responding to publicised need. Deficiencies in individual methods are being countered by customizations, ensemble and hybrid approaches [Langdon *et al.*, 2003a; Runarsson and Sigurdsson, 2004; Li, *et al.*,2005b; Howley and Madden, 2005; Igel, 2005]. These remain the domain of experts and ease of blending of techniques incorporating multi-objective and constraint-based capabilities is awaited with anticipation.

Acknowledgements

This work was partially funded by EPSRC grant GR/T11234/01. The authors wish to thank GSK colleagues, past and present, for their efforts in expressing the nature of their research.

References

Agrafiotis and Cedeno, 2002. Feature selection for structure-activity correlation using binary particle swarms. Journal of Medicinal Chemistry, 45(5):1098-1107.

Amboise and McLachlan 2002. selection bias in gene extraction on the basis of microarray gene-expression data. PNAS, 99(10):6562-6566

Ando and Iba, 2004. Classification of gene expression profile using combinatory method of evolutionary computation and machine learning. GP&EM, 5(2):145-156.

Arimoto and Gifford, 2005. Development of CYP3A4 Inhibition Models: Comparisons of Machine-Learning Techniques and Molecular Descriptors. Journal of Biomolecular Screening, 10(3):197-205

Bains *et al.*, 2004. HERG binding specificity and binding site structure: Evidence from a fragment-based evolutionary computing SAR study. Progress in Biophysics and Molecular Biology, 86(2):205-233.

Banzhaf, *et al.*, 1998. Genetic Programming An Introduction; On the Automatic Evolution of Computer Programs and its Applications; Morgan Kaufmann.

Bao and Sun, 2002. Identifying genes related to drug anticancer mechanisms using support vector machine. FEBS Lett. 521(1-3):109-14.

Barrett, S.J. (2005) INTErSECT "RoCKET" : Robust Classification and Knowledge Engineering Techniques. Presented at : 'Through Collaboration to Innovation', Centre for Advanced Instrumentation Systems, UCL, 16th February 2005.

Bhasin and Raghava, 2004a. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic acids research, 32:W383-W389

Bhasin and Raghava, 2004b. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J. Biological Chemistry, 279(22):23262-23266

Biesheuvel, 2005. Diagnostic Research : improvements in design and analysis. PhD thesis, Universiteit Utrecht, Holland.

Bock and Gough, 2003. Whole-proteome interaction mining. Bioinformatics, 19 (1), 125-135.

Boser *et al.*, 1992. A training algorithm for optimal margin classifiers. 5th Annual ACM Workshop, COLT, 1992

Breiman, 2001. Random forests. Machine Learning, 45:5-32

Breneman 2002. Caco-2 Permeability Modeling: Feature Selection via Sparse Support Vector Machines.Presented at the ADME/Tox symposium at the Orlando ACS meeting,April 2002.

Brown *et al.*, 2000. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc. Natl. Acad. Sci., USA 97:262-267

Burbidge *et al.*, 2001a. STAR Sparsity Through Automated Rejection. In Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence: 6th International Work Conference On Artificial and Natural Neural Networks, IWANN 2001, Proceedings, Part 1, Vol. 2084; Mira, J.; Prieto, A., Eds.; Springer: Granada, Spain, 2001.

Burbidge *et al.*, 2001b. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers in chemistry, 26(1):4-15

Butte, 2002. The use and analysis of microarray data. Nat. Rev. Drug Discov. 1(12):951-60

Byvatov, and Schneider, 2004. SVM-Based Feature Selection for Characterization of Focused Compound Collections. J. Chem. Inf. Comput. Sci., 44(3): 993-999

Byvatov *et al.*, 2005a. From Virtual to Real Screening for D3 Dopamine Receptor Ligands. ChemBioChem, 6(6):997-999

Cedeno and Agrafiotis, 2003. Using particle swarms for the development of QSAR models based on K-nearest neighbor and kernel regression. J.Comput.-Aided Mol. Des.,17:255-263.

Chen, 2004. Support vector machine in chemistry. World Scientific, ISBN 9812389229

Cheng *et al.*, 2004. Insight into the Bioactivity and Metabolism of Human Glucagon Receptor Antagonists from 3D-QSAR Analyses. QSAR & Combinatorial Science, 23(8): 603-620

Congdon and Septor, 2003. Phylogenetic trees using evolutionary search: Initial progress in extending gaphyl to work with genetic data. CEC, pp320-326.

Cristianini and Shawe-Taylor, 2000. An Introduction to support vector machines and other kernel-based learning methods. Cambridge University Press ISBN: 0 521 78019 5

Deutsch, 2003. Evolutionary algorithms for finding optimal gene sets in microarray prediction. Bioinformatics, 19(1):45-52.

Dobson & Doig 2005.Predicting enzyme class from protein structure without alignments. J.Mol.Biol.,345:187-199

Doniger *et al.*, 2002. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. J. of Computational Biol., 9(6): 849-864

Dubey *et al.*, 2005. Support vector machines for learning to identify the critical positions of a protein. Journal of Theoretical Biology, 234(3):351-361

Fradkin, 2005. SVM in Analysis of Cross-Sectional Epidemiological Data. http://dimacs. rutgers. edu/SpecialYears/2002_Epid/EpidSeminarSlides/fradkin.pdf

Eberhart and Hu, 1999. Human tremor analysis using particle swarm optimization. In CEC, pp1927-1930

Eberhart, Kennedy and Shi, 2001, Swarm Intelligence, Morgan Kaufmann.

Fujarewicz and Wiench, 2003. Selecting differentially expressed genes for colon tumor classification. Int. J. Appl. Math. Comput. Sci., 13(3):327-335

Fung and Mangasarian, 2004. A Feature Selection Newton Method for Support Vector Machine Classification. Computational Optimization and Applications 28(2):185-202

Furlanello *et al.*, 2003. Entropy-Based Gene Ranking without Selection Bias for the Predictive Classification of Microarray Data. BMC Bioinformatics, 4:54-74.

Guo *et al.*, 2005. A novel statistical ligand-binding site predictor: application to ATP-binding sites. Protein Engng., Design & Selection, 18(2):65-70

Guyon *et al.*, 2002. Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3):389-422

Hand, 1999. Statistics and data mining: intersecting disciplines.*SIGKDD Explorations*, 1: 16-19

Härdle and Moro, 2004. Survival Analysis with Support vector Machines. Talk at Universite Rene Descartes UFR Biomedicale, Paris http://appel.rz.hu-berlin.de/Zope/ise_stat/wiwi/ise/stat/personen/wh/talks/hae_mor_SVM_%20survival040324.pdf

Heddad *et al.*, 2004. Evolving regular expression-based sequence classifiers for protein nuclear localisation.In: Raidl, *et al.*eds.,Applications of Evolutionary Computing,LNCS 3005, 31-40

Hong and Cho, 2004. Lymphoma cancer classification using genetic programming with SNR features. In Keijzer, *et al.* eds., EuroGP, LNCS 3003, 78-88.

Hou and Xu, 2004. Recent development and application of virtual screening in drug discovery: an overview. Current Pharmaceutical Design, 10: 1011-1033

Howard and Benson, 2003. Evolutionary computation method for pattern recognition of cis-acting sites. Biosystems, 72(1-2):19-27.

Howley and Madden, 2005. The Genetic Kernel Support Vector Machine: Description and Evaluation". Artificial Intelligence Review, to appear.

Huang and Chen, 2005. Support vector machines in sonography: Application to decision making in the diagnosis of breast cancer. Clinical Imaging, 29(3):179-184

Igel, 2005. Multiobjective Model Selection for Support Vector Machines. In C. A. Coello Coello, E. Zitzler, and A. Hernandez Aguirre, editors, Proc. of the Third International Conference on Evolutionary Multi-Criterion Optimization (EMO 2005), LNCS 3410: 534-546

Jerebko, *et al.*, 2005. Support vector machines committee classification method for computer-aided polyp detection in CT colonography. Acad. Radiol., 12(4): 479-486.

Johnson *et al.*, 2003. Metabolic fingerprinting of salt-stressed tomatoes. Phytochemistry, 62(6): 919-928.

Jones, 1999. Genetic and evolutionary algorithms, in: Encyclopedia of Computational Chemistry, Wiley.

Jong *et al.*, 2004. Analysis of Proteomic Pattern Data for Cancer Detection. In Applications of Evolutionary Computing. EvoBIO: Evolutionary Computation and Bioinformatics. Springer, 2004. LNCS, 3005: 41-51

Jorissen and Gilson, 2005. Virtual Screening of Molecular Databases Using a Support Vector Machine. J. Chem. Inf. Model, 45(3): 549-561

Kell, 2002. Defence against the flood. Bioinformatics World, pp16-18.

Kim *et al.*, 2004. Prediction of phosphorylation sites using SVMs. Bioinformatics, 20: 3179-3184.

Kless and Eitrich, 2004. Cytochrome P450 Classification of Drugs with Support Vector Machines Implementing the Nearest Point Algorithm. LNAI, 3303:191-205

Koza, 1992.Genetic Programming: On the Programming of Computers by Means of Natural Selection; MIT Press

Koza *et al.*, 2001. Reverse engineering of metabolic pathways from observed data using genetic programming. Pac. Symp. Biocomp, 2001, 434-435.

Langdon and Barrett, 2004. Genetic programming in data mining for drug discovery. In Ghosh and Jain, eds., Evolutionary Computing in Data Mining, pp211-235. Springer.

Langdon *et al.*, 2001. Genetic programming for combining neural networks for drug discovery. In Roy, *et al.* eds., Soft Computing and Industry Recent Applications, 597-608. Springer. Published 2002.

Langdon *et al.*, 2002. Combining decision trees and neural networks for drug discovery. In Foster, *et al.* eds., EuroGP, LNCS 2278, 60-70.

Langdon *et al.*, 2003a. Comparison of AdaBoost and genetic programming for combining neural networks for drug discovery. In Raidl, *et al.* eds., Applications of Evolutionary Computing, LNCS 2611, pp87-98.

Li *et al.*, 2005. Degree prediction of malignancy in brain glioma using support vector machines. Computers in Biology and Medicine, In Press.

Li *et al.*, 2005b. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1):16-23.

Lin *et al.*, 2005. Piecewise hypersphere modeling by particle swarm optimization in QSAR studies of bioactivities of chemical compounds. J. Chem. Inf. Model., 45(3):535-541.

Listgarten *et al.*, 2004. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clin. Cancer Res., 10: 2725-2737

Liu *et al.*, 2004. QSAR and classification models of a novel series of COX-2 selective inhibitors: 1, 5-diarylimidazoles based on support vector machines. Journal of Computer-Aided Molecular Design 18(6): 389-399

Liu *et al.*, 2005. Preclinical *in vitro* screening assays for drug-like properties. Drug Discovery Today: Technologies, 2(2):179-185

Lu *et al.*, 2004. QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. J. Pharm. Biomed. Anal., 35:679-687.

Malossini et al., 2004. Assessment of SVM reliability for microarrays data analysis. In: proc. 2nd European Workshop on data mining and text mining for bioinformatics, Pisa, Italy, Sept. 2004.

Merkwirth *et al.*, 2004. Ensemble Methods for Classification in Cheminformatics. J. Chem. Inf. Comput. Sci., 44(6): 1971-1978

Miwakeichi *et al.*, 2001. A comparison of non-linear non-parametric models for epilepsy data. Computers in Biology and Medicine, 31(1): 41-57

Moore and Hahn, 2004. An improved grammatical evolution strategy for hierarchical petri net modeling of complex genetic systems. In Raidl, *et al.* eds., Applications of Evolutionary Computing, LNCS 3005, pp63-72.

Moore *et al.*, 2002. Symbolic discriminant analysis of microarray data in automimmune disease. Genetic Epidemiology, 23:57-69.

Muchnik, 2004. Influences on Breast Cancer Survival via SVM Classification in the SEER Database. http://dimacs.rutgers.edu/Events/2004/abstracts/muchnik.html

Ng, 2004. Drugs–From Discovery to Approval. Wiley, New Jersey. ISBN: 0-471-60150-0

Nicolotti *et al.*, 2002. Multiob jective optimization in quantitative structure-activity relationships: Deriving accurate and interpretable QSARs. Journal of Medicinal Chemistry, 45(23):5069-5080.

Norinder, 2003. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. Neurocomputing, 55(1-2): 337-346

Ooi and Tan, 2003. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. Bioinformatics, 19(1):37-44.

Prados *et al.*, 2004. Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents Proteomics, 4(8): 2320-2332

Ratti and Trist, 2001. Continuing evolution of the drug discovery process in the pharmaceutical industry. Pure Appl. Chem.. 73(1):67–75

Reif *et al.*, 2004. Integrated analysis of genetic, genomic, and proteomic data. Expert Review of Proteomics, 1(1):67-75.

Roses, 2002. Genome-based pharmacogenetics and the pharmaceutical industry. Nat. Rev. Drug Discov.1(7):541-9

Runarsson and Sigurdsson, 2004. Asynchronous parallel evolutionary model selection for support vector machines. Neural Information Processing – Lett. & Reviews, 3(3):59-67

Saigo *et al.*, 2004. Protein homology detection using string alignment kernels Bioinformatics, 20: 1682-1689.

Schneider and Fechner, 2004. Advances in the prediction of protein targeting signals Proteomics, 4(6): 1571-1580

Schneider & Fechner, 2005. Computer-based *de novo* design of drug-like molecules. Nat. Rev. Drug Discovery, 4(8):649-663

Schrattenholz,2004. Proteomics: how to control highly dynamic patterns of millions of molecules and interpret changes correctly? Drug Discovery Today: Technologies, 1(1): 1-8

Sebag et al., 2004. ROC-based Evolutionary Learning: Application to Medical Data Mining. Artificial Evolution '03, 384-396 Springer-verlag, LNCS

Seike, *et al.*, 2004. Proteomic signature of human cancer cells. Proteomics, 4(9): 2776-2788

Shawe-Taylor and Cristianini, 2000. An introduction to support vector machines. CUP.

Shen *et al.*, 2004. Hybridized particle swarm algorithm for adaptive structure training of multilayer feed-forward neural network: QSAR studies of bioactivity of organic compounds. Journal of Computational Chemistry, 25:1726-1735.

Shyu *et al.*, 2004. Multiple sequence alignment with evolutionary computation. GP&EM, 5(2):121-144.

Simek *et al.*, 2004. Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. Engineering Applications of Artificial Intelligence, 17: 417-427

Smits *et al.*, 2005. Variable selection in industrial datasets using pareto genetic programming. In Yu, *et al.* eds., Genetic Programming Theory and Practice III. Kluwer.

Solmajer and Zupan, 2004. Optimisation algorithms and natural computing in drug discovery. Drug Discovery Today: Technologies, 1(3): 247-252

Suwa *et al.*, 2004. GPCR and G-protein Coupling Selectivity Prediction Based on SVM with Physico-Chemical Parameters. GIW 2004 Poster Abstract: P056. http://www.jsbi.org/journal/GIW04/GIW04Poster.html

Takahashi *et al.*, 2005. Identification of Dopamine D1 Receptor Agonists and Antagonists under Existing Noise Compounds by TFS-based ANN and SVM. J. Comput. Chem. Jpn., 4(2): 43–48

Takaoka *et al.*, 2003. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. J. Chem. Inf. Comput. Sci., 43(4): 1269-1275.

Teramoto *et al.*, 2005. Prediction of siRNA functionality using generalized string kernel and support vector machine. FEBS Lett. 579(13):2878-82

Thukral *et al.*, 2005. Prediction of Nephrotoxicant Action and Identification of Candidate Toxicity-Related Biomarkers. Toxicologic Pathology, 33(3): 343-355

Tobita *et al.*, 2005. A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors Bioorganic & Medicinal Chemistry Letters, 15:2886-2890

Vinayagam *et al.*, 2004. Appplying support vector machines for gene ontology based gene function prediction. BMC Bioinformatics. 5:116-129

Tsai and Wang, 2005. Evolutionary optimization with data collocation for reverse engineering of biological networks. Bioinformatics, 21(7):1180-1188.

Vapnik, V. N. The Nature of Statistical Learning Theory; Springer: New York, 1995.

Wachowiak *et al.*, 2004. An approach to multimodal biomedical image registration utilizing particle swarm optimization. IEEE Trans on EC, 8(3):289-301.

Wang *et al.*, 2004. Particle swarm optimization and neural network application for QSAR. In HiCOMB.

Wang *et al.*, 2005. Gene selection from microarray data for cancer classification - a machine learning approach. Computational Biology and Chemistry, 29(1): 37-46

Warmuth *et al.*, 2003. Active Learning with Support Vector Machines in the Drug Discovery Process. J. Chem. Inf. Comput. Sci., 43(2): 667-673

Watkins and German, 2002. Metabolomics and biochemical profiling in drug discovery and development. Curr. Opin. Mol. Ther., 4(3): 224-8

Xiao *et al.*, 2003. Gene clustering using self-organizing maps and particle swarm optimization. In HiCOMB

Xu and Hagler 2002. Chemoinformatics and drug discovery. Molecules, 7: 566-600

Xue *et al.*, 2004a. Prediction of P-Glycoprotein Substrates by a Support Vector Machine Approach. J. Chem. Inf. Comput. Sci. 44(4): 1497-1505

Xue, *et al.*, 2004b. QSAR Models for the Prediction of Binding Affinities to Human Serum Albumin Using the Heuristic Method and a Support Vector Machine. J. Chem. Inf. Comput. Sci., 44(5): 1693-1700

Yang and Chou, 2004. Bio-support vector machines for computational proteomics. Bioinformatics, 20: 735 - 741.

Yap and Chen, 2005. Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines. J. Chem. Inf. Model, To appear.

Yap *et al.*, 2004. Prediction of Torsade-Causing Potential of Drugs by Support Vector Machine Approach. Toxicol. Sci., 79: 170-177

Yoon *et al.*, 2003. Analysis of Multiple Single Nucleotide Polymorphisms of Candidate Genes Related to Coronary Heart Disease Susceptibility by Using Support Vector Machines. Clinical Chemistry and Laboratory Medicine, 41(4): 529-534.

Zhao *et al.*, 2004. Diagnosing anorexia based on partial least squares, back-propagation neural network, and support vector machines. J. Chem. Inf. Sci. 44, 2040-2046.