WSC10: 10th Online World Conference on Soft Computing in Industrial Applications, 19 Sep-7 Oct 2005. *To be published by Springer*.

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

SJ Barrett†, WB 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 encompassing many facets [Ng, 2004]. Starting from therapeutic target identification and bioinformatics study [Whittaker, 2004; Lengauer, 2002], through candidate drug discovery and optimisation; to pre-clinical organism-level evaluations and beyond to extensive clinical trials assessing effectiveness and safety of new medicines.

In recent years, with products of human genome project helping to reveal many new disease targets to which drug treatments might be aimed, all the major pharmaceutical companies have invested heavily in the routine ultra-High Throughput Screening (uHTS) of vast numbers of 'drug-like' molecules guided by cheminformatic investigations [Lipinski,2004; Leeson, *et al*., 2004]. Due to the enormous expense of failures of candidate drugs late in their development, uHTS *in vitro* assays now cover liabilities such as possible side effects [Li, 2005] as well as therapeutic properties. In parallel with this, drug design and optimisation increasingly uses computers within *in silico* (virtual) screening [Hou and Xu, 2004; Klebbe, 2004; Schneider and Fechner, 2005]. 'State-of-the-art' *in vitro* experiments now employ DNA micro-array chips to simultaneously explore the expression of thousands of genes potentially involved in disease, treatment and toxicity [Butte, 2002]. Similar advancements are now becoming possible in proteomics [Schrattenholtz, 2004] and metabolomics [Watkins & German, 2002] providing challenges in understanding metabolic pathways and systems biology. Patient-level genetic and single nuclear polymorphism, SNPs [Roses, 2002], data has become more commonly available supporting conventional observational data in epidemiology, clinical trial treatment re-

sponse and early safety studies that continue as on-going pharmacovigilance [Gould,2003].

The curation and storage of all these individual types of data has become more automated, organised and consistent, providing for greater homogeneity and suitability for exploration. Increasingly, vast 'integrated' research datasets are constructed from larger more inhomogeneous combinations of data, from disparate sources and disciplines, to answer novel lines of inquiry, and for hypothesis generation, possibly not initially envisaged at the time of planning data collection. However, conventional multivariate statistical methods, i.e. principal components analysis and partial least squares, well established against smaller, lower-dimensional datasets, are being stretched. Whilst they remain of great utility and continue to be developed in more scaleable commercial tools, they are inherently linear, tending to render them less suitable toward a plethora of newer, ever more complex problem opportunities. Scientists are thus increasingly using data-mining tools such as recursive partitioning and predictive modelling methods to underpin data exploration, using heavy computation to free-up and save scientist time.

Consequently, evaluation and early uptake of novel predictive modelling approaches continues within pharmaceuticals research. Whilst uses of artificial neural networks and genetic algorithms are well established in older application areas [Jones, 1999; Zupan and Gasteiger, 1999; Solmajer and Zupan, 2004], in non-expert hands these may yield suboptimal solutions presenting difficulties in newer areas, including situations when the form of the solution is unclear. More recent machine learning approaches, offer key advantages over these, and we here illustrate Support Vector Machines, Genetic Programming and Particle Swarm Optimisation. The current state of their pharmaceuticals R&D application is reviewed and their future prospects assessed.

2 Support Vector Machines

The Support Vector Machine (SVM) arose from Prof.V.Vapniks' concepts of structural risk minimisation and statistical learning theory [Vapnik, 1992]. An algorithm based upon these ideas was first presented at COLT-92 [Boser *et al*., 1992], and a support vector classifier (SVC) formulation was first presented by [Vapnik, 1995]. Todays' SVC, is a sophisticated synthesis of artificial neural network perceptron-like hyperplane classifier, backed by a sound theory of learning and convergence. It uses robust linear methods and can apply these within kernel spaces to achieve non-linear classifiers with excellent generalisation characteristics.

The simplest SVCs are 'maximal margin' binary classifiers, placing the optimal separating hyperplane, centrally giving the largest allowable separation between the nearest data points of opposite classes in the training set. They use uniform-class subsets of these points (known as support vectors) to construct respective bounding hyperplanes defining a margin which models the decision surface. In accordance with statistical learning theory, for *bias-variance* trade off in learning, this margin-maximisation is tied to a function-limiting to avoid over-fitting. In achieving this, SVCs are constrained to minimise an estimated upper bound on *expected* (not empiri-

cal) risk, as derived from statistical learning theory, assuming training data is drawn independently and identically distributed from some unknown distribution,

$p(x,y)$: $\{ (x_1 y_1), \ldots, (x_l, y_l) \}$ $x_i \in \Re$, *with class* $y_i \in \{ -1 , +1 \}$.

Linear SVCs use the dot product of pairs of input vectors as a distance measure. SVCs can also learn a linear hyperplane after projection of the input to a higher-dimensional kernel-feature space. For efficiency, data mapping to kernel space is not explicitly made, although a sparse new space is effectively created aiding model construction. Kernel spaces allow decision boundaries of apparently arbitrary shape in the input feature space and provide an opportunity to incorporate domain knowledge, enabling solutions to very complex problems of diverse nature [Shawe-Taylor and Cristianini**,** 2004].

Support Vector Regression (SVR) and SVC models achieve a data compression, comprising a linear combination of mapped training examples, the SV subset, using a discovered weighting of input features. Implementations of SVC and SVR are constructed as Linear Programming (LP) or Quadratic Programming (QP) problems using appropriate solver technology. 'Soft margin' SVMs use error terms to handle constraint violations from data-points lying beyond their class 'margin hyperplane', to enable solutions for noisy, or non-linearly separable data. More specific details can be found in [Cristianini and Shawe-Taylor, 2000].

Pros Sound theory and formalism; use robust linear methods; global optimum for convergence; good accuracy, generalisation and robustness to noise; Few user parameters (regularisation parameter, C; kernel parameters), simplify parameterisation compared to neural nets; implicit feature selection; computationally weakly affected by input dimensionality; sparse solution gives fast prediction; Memory linear in the number of training examples.

Cons Complex operation and model opaque to end user; optimal parameter configuration is data dependent; cannot handle missing data; computational cost quadratic with number of examples; QP implementations restricted to Mercer kernels; effectively non-parametric density estimators giving 'point predictions', with no confidences or distributions generated.

2.1 SVM Applications in Pharmaceuticals Research

2.1.1 SVM in Cheminformatics and Quantitative-Structure Activity Relationship (QSAR) Modelling. The role of cheminformatics in drug discovery has been reviewed by [Xu and Hagler, 2002]. An early task is the creation of virtually respresented molecules' and assessment of their likely suitability for synthesis and viability for use in the body. [Byvatov et al. 2003, 2004; Zernov et al., 2003] studied this 'drug-likeness' and report that SVM predictions were more robust than those from neural networks, whilst [Takaoka et al., 2003] has employed SVC to predict chemists' intuitive assessments. Cheminformatics combines chemical properties and high throughput screening measurements , often against novel targets , in large scale struct-

-ture-activity modelling. Trained classifiers enable 'virtual screening' for discovering molecules with specific therapeutic target affinities from potentially millions of virtual representations. Ranking and simple enrichment of actives are key aspects as is the discovery of correlated descriptors, and [Jorissen and Gilson, 2005] developed SVM-based capability to do this.

Reducing the scale of subsequent 'physical' screening of synthesised molecules and the number of synthesis-biotesting cycles for their improvement is an ideal setting for 'active learning' and [Warmuth et al., 2003] have employed SVC in this context. Finding the bio-active conformations of active molecules is key to understanding their mechanisms of action and thus for improving specificity and selectivity, and [Byvatov et al., 2005] have used SVM methodology toward discovering and observing these molecular pharmacophore patterns. SVM uses in the wider field of chemistry have been covered by [Chen, 2004].

Predicting Activity Toward Therapeutic Targets. G-proteins provide such a key interface to intra-cellular signal transduction that 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 with high degrees of success. [Cheng *et al*., 2004] used an RBF-SVR to predict both antagonist compound metabolism and inhibitory activity toward human glucagon receptor in order to select useful 3-d QSAR features. [Byvatov, *et al*., 2005] employed binary SVC optimised via active learning to enrich dopamine receptor agonists then applied SVR to the enriched set to predict relative activities between D2 and D3 receptors to further identify a subset of compounds with required selectivity. [Takahashi *et al*., 2005] used SVC in a 'one versus the rest' context to successfully predict D1 dopamine receptor agonists, antagonists and inactives. [Burbidge, 2004] applied SVM to a variety of QSAR problems and found good performance can be achieved at the expense of sparsity, i.e. a large number of training points are support vectors. Having many support vectors can severely reduce prediction speed in large-scale virtual screening, however [Burbidge *et al*., 2001a] devised an algorithm to counter this.

Predicting Absorption Distribution Metabolism Excretion Toxic effects (ADMET): Amongst the first to investigate the utility of SVC in QSAR modeling, [Burbidge, *et al*., 2001b] favourably compared SVC to back-propagation and radial-basis function neural networks, C5.0 (boosted C4.5 decision trees) and K-nearest-neighbour classifiers against human blood-brain barrier, human oral bioavailability and protein-binding classification problems. [Brenemann *et al*. 2003] have successfully applied SVM to CaCO2 cell permeability prediction. P-glycoprotein (P-gp) active molecular transport in bacterial cells may act as effective efflux pump for antibiotics which are substrates, resulting in drug resistance. [Xue *et al*., 2004a] used Gaussian SVC Recursive Feature Elimination (SVC-RFE) to predict P-gp substrates with ~80% CV accuracy, outperforming probabilistic neural nets and K-NN. [Xue *et al*., 2004b] used similar approach for predicting human intestinal absorption and serum albumin binding. [Doniger *et al*., 2002] demonstrated performance benefits of RBF-SVC with small C over neural networks in working from a small dataset to predict central nervous system (Blood Brain Barrier) permeability with an accuracy above 80%. In contrast to experience with SVC, over-fitting problems were reported for SVR with high C by

[Norinder, 2003] who overcame them using simplex optimization techniques for parameter and feature selection to achieve good predictors for BBB penetration and human intestinal absortion. [Liu *et al*., 2005] report 73% accurracy predicting human oral drug absorption using carefully tuned Gaussian SVR.

Avoiding Adverse Drug Reactions. [Yap *et al*., 2004] used Gaussian kernel SVC to classify drugs in terms of their potential to cause an adverse drug reaction, *torsade de pointes* (TdP). TdP involves multiple mechanisms and their SVM used linear solvation energy relationship descriptors and was optimized by leave-one-out cross validation. Prediction accuracy on an independent set of molecules was in excess of 90% comparing favourably with that from K-NN, probabilistic ANN and C4.5. Accuracy of prediction of TdP-causing agents was substantially improved by SVM, whilst for non-TdP-causing agents discrimination remained comparable to the results obtained by other methods. [Xue *et al*., 2004b] also used SVC, but with RFE in the prediction of TdP inhibition. Chemical inhibition of *Human Ether-a-go-go* Related Gene (HERG) potassium channel is associated with heart arrhymia which can trigger TdP, and [Tobita, *et al*., 2005] have trained a standard RBF SVC using 2-D and molecular fragment features with thresholded pIC50 values to predict HERG inhibition. They report better then 90% accuracy. Non-Steroidal Anti-Inflammatory Drugs (NSAIDS) reduce inflammation by blocking cyclo-oxygenase (COX) enzymes and selective blocking of the COX-2 form reduces gastro-intestinal side effects associated with treatment. [Liu, *et al*., 2004] employed RBF SVC/SVR to discriminate between COX inhibitors.

Predicting Physical Properties. [Lind & Maltseva, 2003] used molecular fingerprint data in an SVR employing a Tanimoto similarity kernel to estimate the aqueous solubility of a set of organic molecules yielding an accuracy comparable to results from other reported methods with the same dataset.

Metabolism and Toxic Effects. Cytochrome p450 enzymes are important chemical (and drug substrate) metabolisers within the body, and significant drug inhibition of these is to be avoided. [Merkwirth *et al*., 2004] compared ridge regression with individual and ensemble K-NN or SVC in predicting CYP450 3A4 inhibitors using medium sized IC50 data excluding substrate molecules. SVC, although computationally most demanding, achieved peak CV/OOT accuracies >90%, with little difference between individuals and ensembles. [Arimoto & Gifford, 2005] employed a larger thresholded HTS CYP3A4 IC50 dataset (with substrates) achieving more than 80% holdout accuracy. This is comparable with that from recursive partitioning and superior to logistic regression, K-NN and Bayes classifiers. [Kriegl *et al*., 2004] achieved similar (~70%) accuracies for CYP3A4 and CYP2D6 inhibition using RBF SVC and SVR, whilst [Kless & Eitrich, 2004] predicted CYP1A2 inhibition with about 90% leave one out cross validation accuracy using an SVM nearest-point algorithm. Yap & Chen, 2005 have since gained further improvements against CYPs 3A4, 2D6 and 2C9 inhibition using Consensus SVMs. Compared to Bayesian-regularised neural network and partial least squares discriminant analysis methods, [Sorich *et al*., 2003] found SVC best able to classify known substrates and non-substrates of 12 human 'drug metabolising' UDP-glucuronosyltransferase (UGT) isoforms, concluding that SVCs were best delineators of the complexities between chemical structure and glucuronidation ability [Miners *et al*. 2004].

2.1.2 SVM in Bioinformatics. In their review of SVM usage in bioinformatics [Byvatov and Schneider, 2003] outlined the major applications up to that point, so here we will briefly re-visit and present an update.

Gene Expression Micro-Array Data in the Prediction of Disease Traits. As with SNPs data, dimensionality P of this input can be extremely large (10Ks of genes) whilst the number of examples N is relatively small (typically a few 10s to 100s). Whilst it is clear that SVMs are well suited to this kind of situation, [Malossini *et al*., 2004] showed that performance can significantly degrade if some training examples are incorrectly labelled. Furthermore increasing the number of correctly labelled training examples does *not* counter the presence of incorrectly labelled examples. Large numbers of poorly correlated, correlated and irrelevant genes also diminish performance, making feature selection essential, and it was for this that [Guyon *et al*., 2002] invented Recursive Feature Elimination (RFE), employing SVC within a wrapper-based approach. [Furlanello *et al*., 2003] developed a faster, 'entropic' form of this eliminating groups of uninteresting genes (rather than one) at a time, whilst [Fujarewicz and Wiench, 2003] devised a heuristic SVC-based Recursive Feature Replacement (RFR) approach. RFR and RFE 'distinctly outperformed' all conventional methods and [Simek *et al*., 2004] found the former best for smaller gene subsets. [Ambroise and Mclachan, 2002] have, however, reported SVM-RFE gene selection bias. [Fung and Mangasarian, 2004] have since reported a fast linear programming SVC handling vast inputs and yet outputting models using very few features. Since their initial usage in this context [Furey *et al*., 2000], SVCs continue to be heavily used to successfully predict cancer cases using case-control gene expression training data for example, [Kun *et al*.,2003; Jarzab *et al*., 2005; Wang *et al*., 2005]. Chemogenomic studies (of functional relationships between genes and drugs) are also increasing, for example, [Bao and Sun, 2002] used multiclass SVC to identify genes related to previously identified anti-cancer drug mechanisms and [Thukral *et al*., 2005] identified drug nephrotoxicity-related gene biomarkers.

Receptor Classification and Protein Function Annotation. SVC prediction of the functional classes of proteins from sequence data is now quite common, and [Karchin *et al*., 2002] were first to achieve this for GPCR families and sub-families using efficient hierarchical multi-class SVC tree (for a comparative evaluation of SVC multi-category methods see [Statnikov *et al*., 2005]). More recently, [Bhasin and Raghava, 2004a] trained 20 SVCs to differentiate GPCR from non-GPCR (99.5% accurracy) and classify to GPCR family (~91% accurracy) and sub-family (96% accurracy). They also report 96% accurracy predicting nuclear receptor sub-family membership [Bhasin and Raghava, 2004b]. [Cai and Lin, 2003] have successfully applied SVC to predicting nucleic-acid binding proteins from their amino acid sequence. [Dobson and Doig, 2005] developed SVM predictor of enzyme classes using simple structural attributes, without sequence alignments.

Gene Functional Classes and Annotation. Since [Brown *et al*., 2000] first employed SVC to predict functional classes of genes, others have continued this. [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. Apparently using default parameterisations, [Gay *et al*., 2002] found a variety of machine learning techniques to achieve a similar level performance for MALDI-TOF Mass Spectrometry (MS) peak intensity prediction, preferring C4.5 and a regression approach. [Jong *et al*., 2004] examined the predictability of cancers using SELDI-TOF mass spectronomy measurements from benign, prostate and ovarian cancer samples, achieving excellent specificity and sensitivity for ovarian cancer with linear SVC. [Seike *et al*., 2004] used SVC within a methodology to rank protein spots in expression profiles from 2D-DIGE (gel electrophoresis) in terms of their discrimatory ability for human cancers. [Prados *et al*., 2004] found SVC to out-perform K-NN, MLP decision tree approaches toward predicting ischemic and haemorrhagic stroke from 42 specimen SELDI-MS data and they applied linear-SVM weight interrogation to further identify a subset candidate biomarkers. [Bock and Gough, 2003] used SVC as an integral part of a learning system that generates protein-protein interaction hypotheses, enabling the development of hypothesised protein interaction networks for bacterial 'design organisms'.

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. For example, sites of: phosphorylation [Kim *et al*., 2004], ATP-binding [Guo, *et al*., 2005], catalysis [Dubey *et al*., 2005] and folding [Han *et al*., 2005]. [Yang and Chou, 2004] achieved improved performance predicting protein cleavage sites by substituting a amino acid similarity matrix for the kernel function, following in a history of kernel modifications in bioinformatics, for example string alignment (in protein homology, [Saigo *et al*., 2004]) and generalised string alignment kernels (in siRNA design for 'gene-silencing', [Teramoto *et al*., 2005]).

2.1.3 SVM in Clinical Diagnosis and Epidemiology

Molecular Genetic Epidemiology. Single-Nucleotide Polymorphisms (SNPs) are common individual base changes within human DNA. Millions of SNPs have been identified. Unlike gene expression measures, SNPs represent unchanging patient-specific variation that may relate to an individuals' prognosis. [Kim and Kim, 2001a,b] were amongst the first to propose an SVM methodology to predict disease using multiple SNP variations. Applying difference scoring to construct input vectors for disease cases and controls by comparison to 'averaged' controls, they they demonstrated feasibility of using SVC. [Yoon *et al*., 2003] developed this, adjusting difference scores at each SNP location by applying weights of chi values from chi-squared test of control allele frequency. They predicted coronary heart disease using SNPs from 10 genes associated with coronary heart disease using polynomial SVC with modest results. They conclude that haplotype data would have been better. [Cohen *et al*., 2003] achieved similar peak accuracies of ~60% trying to predict a high ratio of low-density lipoprotein to high-density lipoprotein using SNPs. [Schwender *et al*., 2004] faired no better with genotypically coded SNPs to predict breast cancer, using imputation to overcome the (common) missing data problem. However, removing examples with missing data, [Listgarten *et al*., 2004] used SNPs from genes of potential relevance to breast cancer, and found that quadratic kernel SVCs (~70% ac-

curracy) out-performed a variety of other methods in prediction of breast cancer, identifying a subset of SNPs that best differentiated cases from controls. Comparing methods to predict SNP effect on protein function, [Krishnan and Westhead, 2003] report competitive results from polynomial SVC but favoured interpretable decision trees with prediction confidences. [Barrett, 2005] summarised the use of normalised binary dominant/recessively coded SNPs. He found SNPs associated with drug effect via iterative training and SNP-removal using 1-norm linear SVC weight-vector interrogation. [Zhang *et al*., 2005] demonstrated successful SNP genotype auto-calling (data generation) using multiclass linear SVC for pre-processed data from a multiplex PCR-microarray system.

Epidemiology. 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] using data covering 112 variables for over 67,000 breast cancer cases from the SEER database, computed multiple SVC models using variable perturbation to generate variable influence estimates to identify 40 'candidate epidemiological factors' with significant influence on the survival time.

Clinical Diagnostics. [Härdle and Moro, 2004] used SVM to model survival analysis.They employed an anisotropic gaussian kernel to predict breast cancer survival with respect to two variables across time. [Zhao *et al*., 2004] used SVC to differentiate anorexic from non-anorexic patients, based upon hair trace element data. [Kim and Kim, 2002] have used patient routine check-up and medical laboratory data to simulate predictions of solidified breast, fatty liver and gastritis endpoints, outlining how SVC could be applied toward case-control data. There is currently a much wider use of SVC in clinical diagnostic areas where large complex data arises from sophisticated equipment such as EEG (epilepsy: [Miwakeichi *et al*., 2001]; depression: [Kalatzis *et al*., 2004]), CT (colon cancer: [Jerebko, *et al*., 2005]), MRI (brain glioma: [Li *et al*., 2005]) and sonography (breast cancer: [Huang & Chen, 2005]). [Majumder *et al*., 2005] compared RBF SVC with RBF Relevance Vector Machines for the prediction of early human oral cancers. Accuracies were very similar overall, although the Bayesian framework of the RVM provides a posterior-probability for classifications.

3 Evolutionary Computing

In contrast to the rigorous mathematical approach of SVMs, evolutionary computation (of which genetic programming is the most advanced variant) appeals to metaphor. The basic idea is to use the ideas of Darwinian evolution within the computer. So we have a population of individuals. A fitness function calculates how good each member of the population is. The better ones are selected to be parents for the next generation. Children are created by crossover and/or mutation of the selected individuals from the

previous generation. As in natural evolution, the children are not identical to their parents. Some are better, some are worse. So in the next generation, selection will again only allow the better individuals to pass their genes onto the next generation. Hopefully overtime and successive generations the population will improved until an individual with satisfactory performance is found.

Such an elegant idea has occurred, apparently independently, to many computer scientists. So who was first is somewhat controversial. However Turing, Rechenberg, Holland and Fogel all make a claim for primacy. From its diverse starting points several subfields of evolutionary computation (Evolution strategy, genetic algorithms, evolutionary programming, etc.) have thrived. Today, even some 40 years after the first papers where published, the field likes to think of itself as a new technique. However, because of its simple appeal, it has been successfully applied many times. Examples include: optimisation, particularly of engineering design, scheduling, economic and financial modelling, fraud detection and data mining. Each sub-field lays stress on different aspects of evolution, e.g. crossover versus mutation, large or small populations and should we represent numbers as bits or as floating point numbers. We will concentrate upon a relative new comer, genetic programming.

3.1 Genetic Programming

Genetic programming, uses Holland's crossover heavy Genetic Algorithm, to evolve programs [Koza 1992; Banzhaf, *et al*., 1998;,Langdon,1998; Langdon and Poli, 2002]. So while other approaches require the software engineer to design an evolutionary friendly way of representing their problem solution, GP does not force this representation to be fixed up front, instead it too can evolve.

Pros Genetic programming combines a flexible problem representation with a powerful search mechanism. Many computational chemistry problems can be expressed as the problem of finding a computer program. E.g., given known properties of a chemical, can we predict some other property (particularly disease binding, toxicology, blood take up). Having recast the problem, the genetic algorithm (GA, used by GP) is a powerful way of searching for a solution which requires minimal assumptions.

Cons Genetic programming offers no guarantee that it will find a suitable solution within an acceptable amount of time. In practise GP has solved difficult but economically interesting problems (for which it is known that no guarantee is possible). While many of the new techniques require more computation time, computer power is increasingly available. Indeed [Buontempo *et al*., 2005] and [Deutsch, 2003, page 49] shows conventional techniques can be out performed in a few minutes. However, from a commercial perspective, spending computer hours (e.g. over night) rather than man-days is a bargain [Bains *et al.,* 2004].

3.1.1 Drug Research Applications of Genetic Programming. In most Pharmaceutical applications, the evolved programs are models. That is, while we can view them as programs which we run and which produce answers, mostly GP is restricted to pro-

ducing functions. These take known facts or measurements (e.g. number of positively charged ions, presence of aromatic rings, acidity) and produce a single number. Then we treat the number as a prediction. For example, a positive number might indicate that the evolved model predicts that the molecule inhibits normal enzyme activity. There is an increasing body of work using evolutionary computation in Biology. For example there are now at least two annual workshops. BioGEC (2002-05) is held in conjunction with the GECCO conference and EvoBIO (2003-05) which is co-sited with EuroGP. Genetic programming figures heavily in both. The June 2004 special issue of the GP journal featured biological applications.

GP in Cheminformatics and QSAR. Genetic programming has been used for combinatorial design [Nicolotti *et al*., 2002], modelling drug bioavailabity [Langdon *et al*., 2002; Langdon *et al*., 2003b; Langdon and Barrett, 2004], and GP ensembles of ANNs have been developed to predict p450 inhibition [Langdon *et al.,* 2001; Langdon *et al*., 2003a].

GP in Bioinformatics. Hot topics include: sequence alignment (typically of either DNA or proteins) [Shyu *et al*., 2004]; protein localisation [Heddad *et al*., 2004; Langdon and Banzhaf, 2005]; using genetic algorithms etc. to infer phylogenetics trees [Congdon and Septor, 2003]; classification and prediction [Hong and Cho, 2004]; recognising parts of proteins (e.g. transmembrane regions [Koza and Andre, 1996]); or in the case of DNA, creating algorithms to find promoters [Howard and Benson, 2003] and other gene regulatory sites. Infrared spectroscopy (wave number), DNA chip and Single Nucleotide Polymorphisms (SNPs) [Reif *et al*., 2004] datasets are noted for having huge numbers of input features. In these cases, while a predictive model might be of use, the immediate problem is to discover which of the thousands of data actually relate to the underlying biology. For example in [Johnson *et al*., 2003], isolation of the relevant wave numbers using GP, revealed new insights into commercial crops. GP based prediction has also been used with DNA chip data in a mode in which, although it generates predictive models, the principle interest is to use GP to sift hundreds or thousands of inputs in order 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, where the bit string GA selects which genes can be used by the second classifier. It is no wonder that GP is increasingly being used in Bioinformatics data mining [Kell, 2002] and increasingly this includes: modelling genetic interactions [Moore and Hahn, 2004] and organisms; inferring metabolic pathways [Koza *et al*., 2001; Tsai and Wang, 2005]; and gene regulatory networks

GP in Clinical Diagnosis and Epidemiology Research. GP has to date made much less overall impact in clinical and epidemiologic areas, although [Biesheuvel, 2005] has applied it in diagnosing patients suspected of pulmonary embolism.

3.2 Particle Swarm Optimisation

Particle Swarm Optimization (PSO) is a population based stochastic optimisation method inspired by observation of swarms of insects, shoals of fish, etc [Eberhart and Kennedy, 1995]. For example, millions of insects can build complex cathedral termite mounds, apparently without central or hierarchical control. Instead each individual acts by themselves in response its environment. Chemical signals provide simple distributed communication between nearby (in space and time) agents. PSO simplifies still further swarms for use in computers for optimisation. The agents are abstracted to particles (like electrons, protons etc., from Physics). These have position and speed. They interact with each other via spring like forces. The particles fly over the problem space. Each time step they sample where they are to determine how good it is. If it is better than any place they have visited, an attractive force is set up which attracts them back to it. There is a similar social cognitive force which attracts the particle to the best place found by the particle's neighbours. A binary extension of PSO (BPSO) is made by replacing the continuous search space by a probability space, i.e. 0..1 in each dimension. At each time step the particle's location is probabilistically converted to a binary string. E.g. a particle at 0.94 along a particular dimension of the problem, has a 94% chance of sampling binary value 1 (true) and only a 6% chance of sampling false.

Pros PSO and BPSO are capable of solving a wide range of very different applications without expensive human up front design.

Cons Like every blind (i.e. problem independent) search technique, PSO do not have a guarantee of success. Nevertheless, as we shall see, despite being originally designed for classic optimisation benchmarks, PSO have been successfully transfered to biological applications.

3.2.1 Biological Applications of Particle Swarm Optimisation. Unlike genetic programming, at present, the use of Particle Swarm Optimisation (PSO) in pharmaceutical research is relatively unexplored. However it is common to use PSO in conjunction with other approaches. This hybrid approach comes from the fact that PSOs naturally search extensively, making them suitable for finding good regions. Often, currently, a more exploitive local method is needed to refine the good starting points found by PSOs into excellent solutions. However as PSOs and their features such as friction (constriction) and momentum become better understood, we anticipate PSOs will tend to be used in a more dominant role.

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.) Some existing drugs (e.g. Aspirin) bind to both COX enzymes, leading to potentially fatal side-effects. [Lu *et al.,* 2004] produce models which can potentially differentiate between binding to the two enzymes by virtual chemicals, i.e.

as an aid to *in silico* design of drugs before the decision is made to manufacture and test the physical chemicals. Both [Wang *et al.,* 2004] and [Shen *et al.,* 2004] use feed-forward artificial neural networks to classify the Bio-activity of chemicals using a few (3-6) features selected by a BPSO. [Wang *et al*., 2004] investigates two ways of using PSO to train the ANN. Either the network is trained in a conventional way or by using another PSO. In [Shen *et al*., 2004] 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 have similar performance at predicting which chemicals will be carcinogenic. The datasets cover typically only cover a few (31-256) chemicals but a large (27-428) number of features are computed for each from the chemical's 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 consider the possibility of over fitting, e.g. by the use of "leave-n-out" cross-validation.

PSO in Bioinformatics. The problem of small but "wide" datasets becomes even more apparent when dealing with DNAchip datasets. [Xiao *et al.,* 2003] suggests a novel combination of PSO and self organising maps (SOM). Instead of finding the few relevant genes, they use SOM to pick clusters of similarly behaved genes from datasets with thousands of gene measurements. The PSO swarm is seeded with crude results produced by the SOM and then used to refine the clusters.

PSO in Clinical Diagnosis and Epidemiology Research. We continue the theme of increasing data width. Two and three dimensional medical images, such as X-Rays and MRI, can contain millions of data per subject. Fortunately the data are regularly arrayed. [Wachowiak *et al.,* 2004] propose a hybrid PSO to solve the computationally demanding task of matching 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 local search within a particle swarm approach. [Eberhart and Hu, 1999] uses a PSO to train an ANN which, using wrist accelerometer data, classifies essential tremor and Parkinson's disease sufferers from control subjects.

4 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 reported machine learning approaches are actually deployed within pharma R&D, so their importance here 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.

Conventional statistical methods are currently better known and understood by scientists. They benefit from their traditional supporting design of experiments, data capture and preparation making them difficult to displace on a wider scale. Statisticians continue to dominate pharmaceutical company quantitative analysis groups. However

statistics is becoming increasingly computational and recognising alternative approaches [Hand, 1999; Breiman, 1996, 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 and from compromises due to cost and lack of prior knowledge. An important recent problem is the integrated analysis of combined 'omics-type data in surrogate biomarker and systems biology research. Here the numerical dominance of variables from genomics, currently swamps those from other types of data in existing methods where all variables (as opposed to the fundamentally different types of information) are treated equally. As individual methods and accompanying validation procedures may only partly cope with problems, multiple methods are often used for comparative analyses in the hope that inappropriate model biases, costly false negatives or effort-producing false positives, are minimised.

SVMs have, however, proved their worth in many areas, and for this technology to make further applications advances there is a need for easier derivation of problem-specific kernel representations, i.e. using structured (ontological) data, or kernel-based data-fusion [Lanckriet *et al.*, 2004]; adequate ways of handling missing data; more widespread generation of confidence measures of prediction and attention to statistical power of datasets in model selection, which itself continues to present problems , especially for SVR end-users. Similar kinds of difficulties hamper the uptake of evolutionary methods by non-expert users, although model transparency (as well as performance) here is a strongly recognized benefit and worthy commercially available tools are now appearing.

Encouragingly, the machine learning research community keeps aware and responds to publicised needs. Deficiencies in individual methods are being countered by customizations, ensemble and hybrid approaches. For example, in QSAR, individual classifiers can be inadequate in the face of vast molecular spaces and multi-mechanism problems. GP classifier fusion was developed to form good ensembles of "weak" or niche classifiers [Langdon and Buxton, 2001a,b,c] using Receiver Operating Characteristics (ROC) curve area as fitness. Whilst GAs are commonly used as feature selectors for SVM they are becoming integrated [Li, *et al.*,2005], and sophisticated hybrids of complementary evolutionary and SVM technique are appearing for kernel development [Howley and Madden, 2005], parameter tuning [Friedrichs and Igel, 2004] , alternative QP solvers [Paquet and Englebrecht, 2003] and model selection [Runarsson and Sigurdsson, 2004; Igel, 2005]. [Huang *et al.,* 2005] used SVM concepts in GA fitness.

An ease of blending of these and other techniques incorporating multi-objective capabilities is awaited with anticipation for challenges in areas like gene regulatory mechanisms discovery [Burckin *et al.,* 2005], selectively non-selective drug design [Roth *et al.*, 2004], clinical trials simulation and 'personalised' of medicines [Bracco, 2002].

5 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.

6 References

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

Ahn *et al*., 2001. A note on applications of support vector machine. http://www.eden.rutgers.edu/~genekim/note.pdf

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 *et al*., 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

Bracco, 2002.Pharmacogenomics and personalised medicine.Pharmacogenomics, 3(2): 166-171

Breiman, 1996. Bagging predictors. Machine Learning, 24(2):123-140

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

Bryant *et al*., 2001. Combining inductive logic programming, active learning and robotics to discover the function of genes. Elect. Trans. in AI, 6(12).

Buontempo *et al*., 2005. Genetic programming for the induction of decision trees to model ecotoxicity data. Journal of Chemical Information and Modeling, 45.

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

Burbidge, 2004. Heuristic methods for support vector machines with applications to drug discovery. Ph.D thesis, University College London, London University, UK

Burckin *et a*l., 2005. Exploring functional relationships between components of gene expression machinery. Nature Structural & Molecular Biology, 12(2):175-182

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

Byvatov and Schneider, 2003. Support vector machine applications in bioinformatics. Appl Bioinformatics. 2(2):67-77

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

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*., 2003. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. J. Chem. Inf. Comput. Sci., 43(6):1882-1889.

Byvatov*, et al*., 2005a. Extraction and visualization of pharmacophore models using support-vector-machines. To appear.

Cai and Chou, 2005. Using Functional Domain Composition To Predict Enzyme Family Classes. J. Proteome Res. 4(1); 109-111.

Cai and Lin, 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim Biophys Acta, 1648(1-2):127-33.

Chen *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

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

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

Cohen *et al*., 2003. Can Support Vector Machines Extract Predictive Power From Single Nucleotide Polymorphisms? http://www.dbmi.columbia.edu/homepages/chw9015/Report-compgenom.doc

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

Corne and Marchiori, 2004. eds. EvoBIO, LNCS 3005.

Corney *et al*., 2004. BioRAT: extracting biological information from full-length papers. Bioinformatics, 20(17):3206-3213.

Cranmer and Bowman, 2005. PhysicsGP: A genetic programming approach to event selection. Computer Physics Communications, 167(3):165-176.

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.

Doniger *et al*., 2002. Predicting CNS Permeability of Drug Molecules: Comparison of Neural Network and Support Vector Machine Algorithms. Journal of Computational Biology, 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 Kennedy, 1995. Particle swarm optimization. Proc. IEEE int'l conf. on neural networks,IV:1942-1948

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

Fradkin, 2005. SVM in Analysis of Cross-Sectional Epidemiological Data. DIMACS Computational and Mathematical Epidemiology Seminar, April 2005. http://dimacs.rutgers.edu/SpecialYears/2002_Epid/EpidSeminarSlides/fradkin.pdf

Friedrichs and Igel, 2004. Evolutionary Tuning of Multiple SVM Parameters. Neurocomputing, 64:107-117

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

Furey, *et al*., 2000. Support vector machine classification and validation of cancer tissue sample using microarray expression data. Bioinformatics 16(10):906-914

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

Gay *et al*., 2002. Peptide mass fingerprinting peak intensity prediction: Extracting knowledge from spectra. Proteomics, 2 (10), 1374-1391

Goodacre *et al*., 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. Trends in Biotechnology, 22(5).

Gould, 2003. Practical pharmacovigilance analysis strategies. Pharmacoepidemiology & Drug Safety, 12: 559–574

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

Huang *et al.*, 2005. Obese Gene Subset Selection: The Maximum Margin Criteria in SVM and Genetic Algorithm.Submitted to:5th IEEE Symposium on Bioinformatics & Bioengineering.

I. Kalatzis, N. Piliouras, E. Ventouras, C. C. Papageorgiou, A. D. Rabavilas and D. Cavouras (2004) Design and implementation of an SVM-based computer classification system for discriminating depressive patients from healthy controls using the P600 component of ERP signals. Computer Methods and Programs in Biomedicine, 75(1): 11-22

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

Jarzab *et al*., 2005. Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. Cancer Res. 2005 Feb 15;65(4):1587-97

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

Karchin *et al*., 2002. Classifying G-protein coupled receptors with support vector machines Bioinformatics, 18: 147-159

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

Kell, 2004. Metabolomics and systems biology: making sense of the soup. Current Opinion in Microbiology, 7(3):296-307.

Kim and Kim, 2002a. A novel statistical diagnosis of clinical data. http://xxx.lanl.gov/ftp/cs/papers/0209/0209001.pdf

Kim and Kim, 2002b. Application of Support Vector Machine to detect an association between multiple SNP variations and a disease or trait. DIMACS workshop, Rutgers University: On the Integration of Diverse Biological Data, 2001. http://dimacs.rutgers.edu/Workshops/Integration/abstracts.html#Kim

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

Klebbe, 2004. Lead identification in post-genomics: computers as a complementary alternative. Drug Discovery Today: Technologies, 1(3): 225-215

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

Koza and Andre, 1996. Classifying protein segments as transmembrane domains using architecture-altering operations in genetic programming. In Angeline and Kinnear, Jr., eds., Advances in Genetic Programming 2, 155-176. MIT Press, 1996.

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.

Kriegl *et al*., 2005. Prediction of Human Cytochrome P450 Inhibition Using Support Vector Machines QSAR & Combinatorial Science. 24(4): 491-502

Krishnan and Westhead, 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. Bioinformatics, 19(17): 2199-2209.

Kun *et al*., 2003. Classifying the estrogen receptor status of breast cancers by expression profiles reveals a poor prognosis subpopulation exhibiting high expression of the ERBB2 receptor. Human Molecular Genetics, 12: 3245-3258

Lanckriet *et al.*, 2004. **A** statistical framework for genomic data fusion. Bioinformatics, 20(16):2626-2635

Langdon and Banzhaf, 2005. Repeated sequences in linear genetic programming genomes. Complex Systems. In press.

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 and Buxton, 2001a. Genetic programming for combining classifiers. In Spector, *et al*. eds., GECCO, 66-73,
Langdon and Buxton, 2001b. Genetic programming for improved receiver operating characteristics. In Kittler and Roli, eds., MCS, LNCS 2096, 68-77.
Langdon and Buxton, 2001c. Evolving receiver operating characteristics for data fusion. In Miller, *et al*. eds., EuroGP, LNCS 2038, 87-96.
Langdon and Buxton, 2004. Genetic programming for mining DNA chip data from cancer patients. GP&EM, 5(3):251-257.
Langdon and Poli, 2002. Foundations of Genetic Programming. 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.
Langdon *et al*., 2003b. Predicting biochemical interactions - human P450 2D6 enzyme inhibition. In Sarker, *et al*. eds., CEC.
Langdon, 1998. Genetic Programming and Data Structures, Kluwer.
Leeson *et al*., 2004. Drug-like properties: guiding principles for design – or chemical prejudice? Drug Discovery Today: Technologies, 1(3):189-195
Lengauer, 2002. Bioinformatics. From Genomes to Drugs. Wiley-VCH, Weinheim, Germany, Vols I,II, ISBN: 3-527-29988-2
Li, 2005. Preclinical *in vitro* screening assays for drug-like properties. Drug Discovery Today: Technologies, 2(2):179-185
Li *et al*., 2005. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 85(1):16-23.
Li *et al*., 2005. Degree prediction of malignancy in brain glioma using support vector machines. Computers in Biology and Medicine, In Press.
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.
Lind and Maltseva, 2003. Support Vector Machines for the Estimation of Aqueous Solubility. J. Chem. Inf. Comput. Sci., 43(6): 1855-1859
Lipinski, 2004. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies 1(4):337-341
Listgarten *et al*., 2004. Predictive Models for Breast Cancer Susceptibility from Multiple Single Nucleotide Polymorphisms. Clin. Cancer Res., 10: 2725-2737
Liu *et al*., 2003. Diagnosing Breast Cancer Based on Support Vector Machines. J. Chem. Inf. Comput. Sci., 43(3); 900-907
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
Lu *et al*., 2004. QSAR analysis of cyclooxygenase inhibitor using particle swarm optimization and multiple linear regression. J. Pharm. Biomed. Anal., 35:679-687.
Majumder *et al.* 2005. Relevance vector machine for optical diagnosis of cancer. Lasers in Surgery and Medicine, 36 (4): 323-333
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

Miners *et al*., 2004. Predicting human drug glucuronidation parameter: Application of In Vitro and In Silico Modeling Approaches. Annual Review of Pharmacology and Toxicology, 44: 1-25

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. John Wiley & Sons,Hoboken, 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.

Paquet and Englebrecht, 2003. Training support vector machines with particle swarms. Intl. Conf. Neural Networks, Portland, Oregon.

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

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

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

Roth *et al*., 2004. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. Nature Reviews - Drug Discovery, 3:353-359

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

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

Schwender *et al*., 2004. A pilot study on the application of statistical classification procedures to molecular epidemiological data. Toxicol. Lett., 151: 291-299

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

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

Shawe-Taylor and Cristianini, 2004. Kernel methods for pattern analysis. Cambridge University Press ISBN: 0 521 81397 2

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.

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

Sorich *et al*., 2003. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. J. Chem. Inf. Comput. Sci., 43(6):2019-2024.

Statnikov *et al.,* 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics, 21(5): 631-43

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.

Vapnik, V. Principles of Risk Minimization for Learning Theory. In Advances in Neural Information Processing Systems, Vol. 4; Moody, J. E.; Hanson, S. J.; Lippmann, R. P., Eds.; Morgan Kaufmann Publishers, Inc.: 1992.

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

Whittaker, 2004. The role of bioinformatics in target validation.Drug Discovery Today:Technologies,1(2):125-133

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.

Yao *et al*., 2004. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. J. Chem. Inf. Comput. Sci., 44(4): 1257-1266.

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.

Zernov *et al*., 2003. Drug discovery using support vector machines. The case studies of drug-likeness, agro-chemical-likeness, and enzyme inhibition predictions. J. Chem. Inf. Sci., 43:2048-2056.

Zhang *et al.,* 2005. SNP auto-calling using support vector machines. Proc. of the 6th International Symposium on Computational Biology and Genome Informatics, Salt Lake City, July 21-26, 2005 (to appear).

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.

Zupan and Gasteiger, 1999. Neural Networks in Chemistry and Drug Design: An Introduction, 2nd Edition. John Wiley, ISBN: 3-527-29778-2, 400pp.