ASSESSMENT OF TAXOL ACTIVITY: BIOAVAILABILITY IN HUMAN PHYSIOLOGICAL FLUIDS AND QSAR OF TAXOL ANALOGUES BASED ON A NEURAL NETWORK DESIGN

by

Stanislav Robert Svojanovsky

Ing., University of Pardubice, Czech Republic, 1982

M.S., Western Michigan University, Kalamazoo, MI 49008, 1993

Submitted to the Department of Chemistry and the Faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Date Defended: September 16, 1999

ABSTRACT

Stanislav R. Svojanovsky, September 1999

University of Kansas

Taxol is an anticancer agent that is highly protein bound. This feature is addressed in the first chapter by developing a competitive, enzyme-linked immunosorbent assay (ELISA) to quantify bioavailability of taxol in human fluids: plasma, plasma ultrafiltrate and salivary fluids. An optimization procedure decreased the detection limit to pg/ml. The working range encompasses five orders in magnitude, i.e. from pg/ml to 100 ng/ml. ELISA was then applied to detect taxol in samples from a cancer patient. The results show the taxol concentration in plasma (sub-μg/ml), ultrafiltrate (ng/ml) and salivary fluids (sub-ng/ml).

The second chapter describes a new class of immunosensors in which the binding event is directly transduced, so there is no requirement for a chemical label. The ligand is *Staphylococcal Enterotoxin B* (SEB), while the receptor is the anti-SEB antibody. A high sensitivity radioimmunoassay was developed to detect about 1 pmol/cm^2 of bioactive SEB-antibody, immobilized on the sensor surface. This technology has the potential for a low cost diagnostic assay and could be used in the medical and food industries.

Chapter three describes a back-propagation neural network (BPNN) design for 50 taxol analogues. Initial system consists of 27 calculated structural descriptors, while the outputs are the measured antitumor activities against 4 types of cancer (breast, ovarian, lung and the overall average GI_{50}). The optimization process leads to dimensional reduction with increased accuracy (in comparison to other parametric or non-parametric techniques), in terms of correlation, i.e. 0.831 ovarian, 0.945 lung, 0.913 breast cancer, and 0.886 for the index GI_{50}. Calculated molecular properties and the anticancer activities are the only required input and output variables, so that the potential of the compound can be established prior to the synthesis.

Quantitative structure-activity relationships (QSAR) for a set of 61 taxol analogues previously excluded from a BPNN design are explained in chapter four. The input data are calculated descriptors (lipophilicity, molar refractivity, dipole moment plus dipole vectors, steric plus conformation energies, and heat of formation). High antitumor activity is predicted for taxol analogues with a substituent propionic, 1-methyl-2-pyrrolecarboxylic, or crotonic acid in position C-10.

To my wife Romana

ACKNOWLEDGMENTS

I did not achieve this accomplishment alone. It was a joint effort of many people who helped, supported and guided me to successfully complete this dissertation.

I would like to thank my advisor, Dr. George S. Wilson, for his continuous support throughout my entire graduate career. Under his leadership I was able to gain additional knowledge from different fields of science and apply my training in statistics to the areas of computational chemistry and chemoinformatics. I am thankful for his guidance in allowing me to become an independent scientist and for the opportunity to focus on diverse research disciplines.
Many thanks to all the members of the Wilson group for the helpful discussions and friendship.

I greatly appreciate the help and guidance provided by Dr. Swapan Chakrabarti from the Department of Electrical Engineering and Computer Science (EECS). This collaboration was an excellent learning experience, which enhanced my understanding of the potential of the neural network applications.

I would like to thank my dissertation committee and faculty members of the Chemistry and EECS departments at The University of Kansas for their support.

Finally, I would like to express my sincere gratitude to my wife, her parents and my parents. If it were not for them, none of this would have been possible, so I just want to thank you all for believing in me and giving me your help in countless different ways.

TABLE OF CONTENTS

LIST OF TABLES xii

LIST OF FIGURES xiv

CHAPTER 1 - HIGH SENSITIVITY ELISA DETERMINATION OF TAXOL IN VARIOUS HUMAN BIOLOGICAL FLUIDS 1

ABSTRACT 2

INTRODUCTION 3

EXPERIMENTAL 10

Materials 10

METHODS 11

Synthesis of BSA-7-succinyltaxol conjugate (BSA-7-SucTax) 11

Characterization of taxol antibodies 11

ELISA for taxol detection in physiological fluids 13

Taxol Ab test for taxol analogues 16

RESULTS 17

DISCUSSION 37

REFERENCES 39

CHAPTER 2 - MEASUREMENT OF THE BIOLOGICAL ACTIVITY OF PROTEINS IMMOBILIZED ON SENSORS 45

ABSTRACT 46

INTRODUCTION 47

Staphylococcal enterotoxin B (SEB) 50

EXPERIMENTAL 52

Material 52

Characterization of anti-SEB antiserum (SEB-Ab) 53

Purification of SEB-Ab antiserum 54

Radioiodination procedure 56

Antibody immobilization on the sensor 57

RESULTS AND DISCUSSION 60

CONCLUSION 65

REFERENCES 67

CHAPTER 3 - DESIGN OF NEURAL NETWORK MODELS FOR SCREENING ANTICANCER ACTIVITIES IN TAXOL ANALOGUES 70

ABSTRACT 71

INTRODUCTION 72

COLLECTING AND PREPARING DATA 82

Data requirements 82

Data validation 83

Data partitioning 83

DESIGN, TRAINING AND TESTING OF THE NN PROTOTYPE 84

Data preparation 85

Selection of the training and testing set 86

Selection of NN type and architecture 87

Training the NN prototype 87

Testing the NN prototype 88

Experimental optimization 88

Comparison of the NN performance with other methods 89

COMMON PROBLEMS IN TRAINING AND TESTING THE NN PROTOTYPE 91

Poor training performance 91

Poor generalization performance 91

EXPERIMENTAL DESIGN OF NEURAL NETWORK PROTOTYPES ..93

Collecting data 93

Input data 93

Output data 96

Data preparation 97

Selection of the training and testing set 97

Selection of the NN type and architecture 98

Analysis of the prediction accuracy 100

Dimensionality reduction 101

Selecting the number of feature vectors 106

Selecting other parameters 107

Experimental optimization 110

Comparison of the NN performance with other methods 114

RESULTS 117

CONCLUSION 119

REFERENCES 120

CHAPTER 4 - QSAR OF TAXOL ANALOGUES USING NEURAL NETWORK PROTOTYPES 126

ABSTRACT 127

INTRODUCTION 128

BACKGROUND 132

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) 133

TAXOL ANALOGUES 136

Input data for the neural network design 136

Output data – anticancer activities 137

Assay for determination of anticancer activity 137

NEURAL NETWORK PROTOTYPE 140

Physicochemical descriptors 140

Neural network architecture and accuracy 147

TESTING THE NEURAL NETWORK PROTOTYPE 148

Input data 148

Anticancer activity prediction 148

RESULS AND DISCUSSION 154

CONCLUSION 161

REFERENCES 163

LIST OF TABLES

CHAPTER 1

Table 1. Limit of detection of taxol in human physiological fluids for different matrices 19

Table 2. Taxol recovery from spiked samples of physiological fluids 26

CHAPTER 2

Table 1. Apparent binding constant of SEB-Ab antiserum to SEB 60

Table 2. Radioactivity measurements 63

CHAPTER 3

Table 1. Parameters of the feature vector 95

Table 2. Cell lines used in calculation of the activity average 96

Table 3. Analysis of correlation matrix 102

Table 4. Principal component analysis 104

Table 5. Pattern analysis 105

Table 6. Subset of the input data for 10 compounds 106

Table 7. Optimization the number of feature vectors 107

Table 8. BPNN design for OVA and LNS cancer 108

Table 9. BPNN design for BRE cancer and GI_{50} 109

Table 10. Best BPNN prototypes 110

Table 11. Optimized BPNN for OVA and LNS cancer 111

Table 12. Optimized BPNN for BRE cancer and GI_{50} 112

Table 13. Optimal BPNN prototypes 113

Table 14. Comparison of the BPNN with other methods 115

Table 15. Accuracy of the analog BPNN and multiregression 117

Table 16. Accuracy of the binary BPNN and Bayes' rule 117

Table 17. Correlation between predicted and real activities 118

CHAPTER 4

Table 1. Physicochemical parameters used for the design of the neural network prototype 141

Table 2. Predicted activity against ovarian cancer 149

Table 3. Predicted activity against lung cancer 150

Table 4. Predicted activity against breast cancer 150

Table 5. Predicted activity against the average index GI_{50} 151

Table 6. Analogues with high activities against OVA and LNS 152

Table 7. Analogues with high activities against BRE and GI_{50} 153

Table 8. Ordered taxol analogues with the highest predicted activities for each cancer type 154

Table 9. Taxol analogues with the highest predicted activities 157

Table 10. The structures of the substituent groups 159

LIST OF FIGURES

CHAPTER 1

Figure1. Structural formula of the taxol molecule.....4

Figure 2. Flow chart of ELISA method 14

Figure 3. Titration curve for taxol antibody to BSA-7-SucTax conjugate (Tax-Ab at 10 μg/ml) 18

Figure 4. Optimization process for taxol antibody 20

Figure 5. Optimization process for BSA-7SucTax conjugate 21

Figure 6. Taxol calibration curve in ultrafiltrate matrix (pH 7.0) 23

Figure 7. Taxol calibration curve in saliva matrix (pH 7.0) 24

Figure 8. Taxol calibration curve in pooled plasma matrix (pH 7.0) 25

Figure 9. Detection of taxol in human biological fluids by ELISA method 28

Figure 10. Simulated model with the actual detection of taxol in salivary fluids 30

Figure 11. Simulated model with the actual detection of taxol in ultrafiltrate ...31

Figure 12. Simulated model with the actual detection of taxol in plasma 32

Figure 13. Structures of taxol analogues with high response towards Tax-Ab ..34

Figure 14. Calibration curves for the active taxol analogues 35

CHAPTER 2

Figure 1. Cross-section profile of the immunosensor 49

Figure 2. Immobilization procedure for covalent attachment of SEB-Ab to a silica surface 59

Figure 3. Titration curve for SEB-Ab antiserum to SEB (conc. 5 μg/ml) 61

CHAPTER 3

Figure 1. Formal neuron 76

Figure 2. Classical BPNN architecture 78

Figure 3. Components of the NN design 84

CHAPTER 4

Figure 1. Profile of the training data set (OVA) 145

Figure 2. Profile of the averages for both classes (OVA) 146

Figure 3. Structural formulas of the baccatin III (A), taxol (B) and taxotere (C) molecules 155

Figure 4. Structures of the 7-O and 10-O taxol analogues 156

Chapter 1

HIGH SENSITIVITY ELISA DETERMINATION OF TAXOL IN VARIOUS HUMAN BIOLOGICAL FLUIDS

Abstract

Taxol is a novel agent with high activity in the treatment of patients with several malignant tumors including those resistant to other cytotoxic drugs. The therapeutic index of this promising anticancer drug could be further increased by the exploration of its pharmacokinetic and pharmacodynamic relationship in cancer patients. Since taxol is highly protein bound, a very specific and highly sensitive analytical method is required in order to determine free, protein unbound and biologically active taxol species in human physiological fluids: plasma, plasma ultrafiltrate and salivary fluids. In order to accomplish this, a new, indirect competitive enzyme-linked immunosorbent assay (ELISA), for quantitating such a low bioactive taxol concentration level, has been developed. While employing this technique, after systematic optimization of the experimental conditions, we are able to detect the anticipated taxol in plasma ultrafiltrate and salivary fluids at the concentration level of subpicogram per milliliter. The working range of the assay is approximately five orders in magnitude, i.e. from pg/ml to 100 ng/ml. The assay shows an interaction with some of the taxol analogs as well. The clinical part of this study verified the working range of the ELISA method using samples of physiological fluids from a cancer patient treated with three hours intravenous (IV) infusion of this drug. Our results of taxol determination in plasma, plasma ultrafiltrate and saliva demonstrate the applicability of this assay for further pharmacokinetic studies of free, biologically active taxol species in cancer patients.

Introduction

Taxol (Paclitaxel) is a plant product isolated from the cortex of the Pacific yew (*Taxus brevifolia*). It is one of the most active anticancer agents introduced in clinical oncology practice in the last decade and the first of a new class of cytotoxic drugs with high activity against tumors resistant to other drugs.

Purified taxol is a white crystalline powder with the empirical formula $C_{47}H_{51}NO_{14}$ and a molecular weight of 853.9 daltons.

Taxol molecule is highly lipophilic (LogP around 4) with poor solubility (less than 0.01 mg/ml, i.e. about 1.2×10^{-5} M) in water (Mathew *et al.*, 1992; Kingston *et al.*, 1998), and melts at around 216-217°C. Taxol is a taxane diterpenoid characterized by its taxane ring system with a four-membered oxetane ring and ester side chain at C-13 (the N-benzoyl-3-phenylisoserine ester of the compound *baccatin III*). No X-ray structure of taxol has been reported yet but the X-ray structure of taxotere (an analogue with the N-benzoyl group replaced by a t-butoxycarbonyl group) shows a cup-like shape for the taxane skeleton.

The chemical name for paclitaxel is:

5β, 20-Epoxy-1, 2α, 4, 7β, 10β, 13α-hexahydroxyltax-11-en-9-one 4, 10-diacetate 2-benzoate 13-ester with (2'R, 3'S)-N-benzoyl-3-phenylisoserine.

Figure 1 shows the structural formula of taxol.

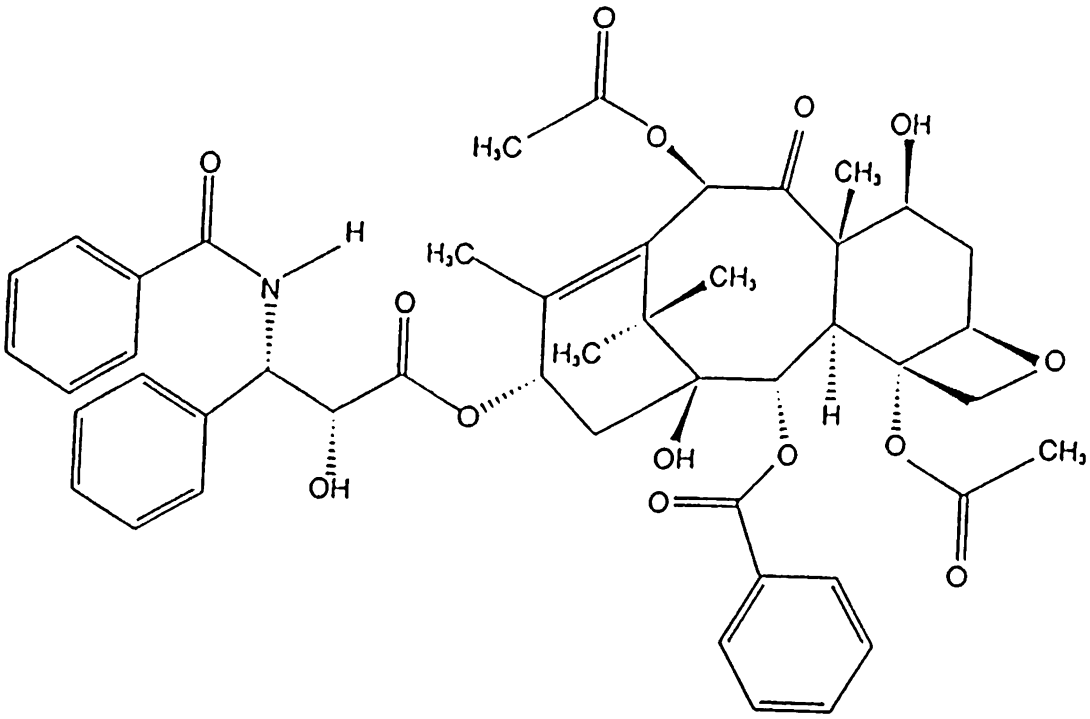


Figure 1. Structural formula of the taxol molecule

As a natural product, taxol was detected as the active component of the bark of the Pacific and Western yew in 1971 by the National Cancer Institute (NCI), Drug Development Program (Wani *et al.*, 1971). Further clinical development of taxol was delayed for many years due to difficulties in large-scale isolation and extraction where the yield of taxol was less than 0.02% from dried bark of *Taxus* species *brevifolia* or *baccata*.

In the late 1970s the mechanism of action of taxol was described as unique among plant alkaloids, making it the prototype for a new class of chemotherapeutic agents. In 1984 investigational new drug (IND) application was approved and clinical trials began. The Johns Hopkins University School of Medicine reported very positive results in the treatment of advanced ovarian cancer in 1989. Bristol-Myers Squibb was then selected by NCI as a partner to commercialize the drug. The Food and Drug Administration (FDA) granted marketing approval for a semi-synthetic form of paclitaxel for treatment of certain cancers of the breast and ovary. Use for the treatment of metastatic breast cancer received marketing approval in 1994. Groups led by K. C. Nicolaou of the Scripps Research Institute and R. Holton of Florida State University announced the complete synthesis of taxol. From 1995, research has focused on additional taxol applications in the area of Kaposi's sarcoma and other kinds of cancer. In 1997, a report of the development of a highly efficient, more water-soluble polymer-taxol conjugate was published (Li *et al.*, 1997). The conjugate was prepared by attaching taxol to poly(L-glutamic acid) (PG). This more water-

soluble polymer-taxol conjugate demonstrates a high antitumor efficacy in two rodent tumor models and is still under further investigation and research.

Taxol affects the stabilization of microtubules, which are intracellular structures vital to mitosis and other critical cell functions (Schiff *et al.*, 1979; Rowinsky *et al.*, 1992). It targets rapidly dividing cancer cells and prevents them from replicating. When cells reproduce using mitosis, the chromosomes, which contain the cells' genetic information, are divided into two analogous sets. Spindle fibers are then produced to guide both parts to the different areas of the cell. The mother cell then splits down the middle and the two daughter cells create their own compliments of chromosomes. Taxol binds to tubulin dimers (and recent experiments provide supporting evidence for binding to the N-terminal amino acids of the β-subunit of tubulin), therefore prevents the mother cell from assembling spindle fibers and that inhibits dynamic cell division. Preclinical studies revealed experimental antitumor activity, including tumors resistant to prior therapy and the risk of developing several hypersensitivity reactions. Taxol is most commonly used in combination with other chemotherapy drugs such as 5-fluorouracil (5-FU), Adriamycin, Vinorelbine, Cytoxan and Cisplatin. The low taxol water solubility makes it very difficult to get into the bloodstream; moreover, it can only be delivered in a toxic solvent. Furthermore, once in the body, it attacks bone marrow, nerve fibers, and mucous membranes, as well as cancer cells. The very low solubility in water raises serious problems related to its formulation. Taxol (currently produced via semi-synthetic process from a precursor that can be obtained in higher yield) is formulated in a vehicle (containing Cremophor EL and

alcohol) as a colorless to slightly yellow, viscous solution, available in single-dose vials containing 30 mg/ml. Cremophor EL, a non-ionic surfactant, is a polyoxyethylated castor oil that is commonly used to dissolve several other water-insoluble drugs, including cyclosporine, and diazepam. This solvent probably causes the allergic reactions observed in some patients. However, results from initial clinical studies disclosed that hypersensitivity reactions could be sufficiently reduced in cancer patients by pretreatment with glucocorticoids (such as dexamethasone) and antihistamines (such as ranitidine or cimetidine). Myelosuppression, particularly neutropenia, has been the major dose-limiting toxicity (Koeller *et al.*, 1994). Several tumors have been found to be highly responsive to taxol treatment, with response rates over 30% for ovarian cancer (Gregory *et al.*, 1993; Ozols, 1995; Hajek *et al.*, 1996), higher than 50% for breast cancer (Kalous *et al.*, 1997), and over 20% for non-small cell lung cancer (Natale, 1995; Brzezny *et al.*, 1997), all resistant to previous chemotherapy. Since taxol is highly (more than 93%) bound to human serum proteins, analytical methods to detect free, unbound and bioactive fractions of taxol have to be very specific and highly sensitive.

Current analytical methods for detection of taxol (taxanes) include high performance liquid chromatography–HPLC (Rizzo *et al.*, 1990; Cardelina 1991; Harvey *et al.*, 1991; Auriola *et al.*, 1992; Willey *et al.*, 1993), multimodal thin layer chromatography-MTLC (Stasko *et al.*, 1989), a tubulin-dependent biochemical assay (Hamel et al., 1982), and micellar electrokinetic chromatography-MEKC (Chan *et al.*,

1994; Hempel *et al.*, 1996). The chromatographic methods for quantitating taxol in different samples containing closely related taxanes (cephalomannine, baccatin III, and 10-deacetylbaccatin III) are relatively insensitive and time consuming and most of all, they cannot detect the biologically active forms of taxol in physiological fluids in ultra-low concentrations, i.e. pg/ml.

This would leave the immunoassay as the preferred method for quantifying analysis of taxol (Jaziri *et al.*, 1991; Leu *et al.*, 1993; Grothaus *et al.*, 1993, 1995) at the required concentration levels. These assays provide highly sensitive and precise methods for the estimation of biological parameters, with the advantage that they can handle a large number of samples, which may be then analyzed more rapidly. Among the various immunoassay formats, the most common is the enzyme-linked immunosorbent assay (ELISA), which includes a separation step to remove the free antibody portion from the bound antibody. Several assay configurations can be used in ELISA, depending on the analyte and the availability and purity of the antibodies and the corresponding enzyme labels. In the 'sandwich-type' ELISA technique a 'capture' antibody is immobilized on a solid support of the microplate, sample is added and the analyte is allowed to bind to the antibody. After a washing step a second enzyme labeled 'detection' antibody is used to detect this complex. The assays are very sensitive because of two analyte-antibody binding steps. The competitive ELISA, used in this study, requires that the antibody compete for the immobilized and free analyte. The detection limit depends upon antibody affinity.

In this study, the results are presented on the development of a sensitive ELISA method for the detection of taxol in physiological fluids, together with a clinical application to determine the taxol level in human physiological samples. Samples were obtained from one cancer patient following administration of taxol by a three-hour intravenous infusion.

Note:

Taxol is used in this monograph to refer to the drug that now has the generic name paclitaxel and the registered trade name TAXOL® (Bristol-Myers Squibb Company, New York, NY).

Taxotere is used in this monograph to refer to the drug that now has the generic name docetaxel and the registered trade name TAXOTERE® (Rhône-Poulenc Rorer Pharmaceuticals Inc., Collegeville, PA).

Experimental

Materials

7-succinyltaxol (7-SucTax) used in this study was purchased from Hawaii Biotechnology Group, Inc. (Aiea, HA) and partially obtained as a kind donation from Prof. Richard R. Himes, Department of Biochemistry, University of Kansas, Lawrence, KS 66045. Taxol (*Taxus brevifolia;* Baccatin III N-benzyl-3-phenylisoserine ester; purity > 99% by HPLC) was purchased from Calbiochem (La Jolla, CA). Anti-taxol monoclonal antibody (Tax-Ab) was purchased from Hawaii Biotechnology Group, Inc. (Aiea, HA). Bovine serum albumin (BSA) and (EDAC) 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide were purchased from Sigma Chemical Co. (St. Louis, MO). N-hydroxysulfosuccinimide (NHS) was purchased from Pierce Chemical Co. (Rockford, IL). Goat anti-mouse horseradish peroxidase (Gt x Mo-HRP) was purchased from Jackson Laboratories (West Grove, PA) and Spectrapor membrane tubing (with molecular cut off 8-20kDa) from Spectrum Medical Industries, Inc. (Los Angeles, CA). Prof. Gunda Georg, Department of Medicinal Chemistry, University of Kansas, Lawrence, KS 66045, donated taxol analogues. All other chemicals were purchased from Fisher Scientific (St. Louis, MO) or Aldrich Chemical Co. (Milwaukee, WI).

Methods

Synthesis of BSA-7-succinyltaxol conjugate (BSA-7-SucTax)

To a 500 µl solution of 7-SucTax (0.6 mg, 0.63 µmol) in anhydrous N,N-dimethyl formamide (DMF) were added 700 µl phosphate buffered saline (PBS, pH 7.0; 0.01 M), 300 µl of a 1.0 M EDAC solution (191.4 mg in 1 ml water) and a solution of NHS (0.2 mg, 0.92 µmol) in 500 µl PBS (pH 7.0; 0.01 M). The reaction was stirred at room temperature for 10 min and then a solution of BSA (4 mg, 60 nmol) in 1.00 ml of PBS (pH 7.0; 0.01 M) was slowly added. After stirring for 24 h at room temperature the dialysis was performed using Spectrapor membrane tubing and the concentration of BSA-7-SucTax protein was calculated on the basis of BSA concentration.

Characterization of taxol antibodies

The term affinity is used to describe the strength of binding between the antibody (Ab) and the antigen (Ag). Electrostatic forces, such as coulombic attractions, hydrogen bonds and hydrophobic interactions generally drive the binding reaction. The non-covalent interactions that result from this affinity are of particular importance in biological processes. High affinity can be seen as binding involving a lot of energy or of longer duration.

The apparent binding (affinity) constant (K_a) is analogous to the equilibrium constant in a chemical reaction, called biospecific interaction, which is characterized by the

relatively strong non-covalent interactions between the Ag and the adequate Ab to form a stable Ag-Ab complex. K_a is a quantitative term described by the following equation:

$$[Ab] + [Ag] \leftrightarrow [Ab:Ag]$$

$$K_a = k_{association} / k_{dissociation} = [Ab:Ag] / [Ab][Ag]$$

Where $k_{association}$ is the kinetic 'on rate' for the forward reaction, while $k_{dissociation}$ is the 'off rate' for the opposite reaction.

A low affinity system ($K \sim 10^6\ M^{-1}$) is used to measure substances at higher concentration, while a higher affinity ($K > 10^8\ M^{-1}$) system allows the measurements of substances at very low concentrations and/or trace analysis.

The binding characteristics of the taxol antibody towards its antigen coated on the microtiter plates (BSA-7-SucTax) were studied. A standard procedure to obtain titration curves (optical density vs. antibody concentration) via direct, noncompetitive ELISA was performed on high-binding polystyrene ELISA 96 well microtiter plates (Corning 25801). The plate wells were coated for 2 h at 37°C with 100 µl per well of BSA-7-SucTax conjugate (5 µg/ml) in sodium bicarbonate (coating) buffer (pH 9.5; 0.1 M). The plates were then washed four times with PBS (pH 7.4; 0.01 M)

containing 0.5% (v/v) Tween 20 (PBS-T-20) and blocked for 1 h at 37°C with 100 µl of blocking buffer (PBS-T-20, containing 0.2% BSA of ELISA grade (98-99%)). After washing the plate four times with wash buffer (PBS-T-20) and drying them at room temperature, taxol antibody at the concentration of 10 µg/ml was added to the first two wells and serially diluted by two wells over the microtiter plate. The last four wells were used as blank to eliminate the background signal. Plates were incubated at 37°C for 1 hr and washed again with wash buffer (pH 7.4; 0.01 M). Then, Gt x Mo-HRP conjugate in blocking buffer (1:15,000 dilution, 100 µl) was uniformly added, the plates incubated for 1 hr at 37°C and washed four times with nanopure water before adding 100 µl of microwell peroxidase substrate (TMB) solution uniformly to each well. The colorimetric reaction was stopped after 10 min by adding 50 µl of 1 N HCl to each well, and the absorbance was measured on Kinetic Vmax Microplate Reader (Molecular Devices Corp., Menlo Park, CA) using a sample wavelength of 450 and 650 nm.

ELISA for taxol detection in physiological fluids

A competitive inhibition ELISA was performed on polystyrene ELISA 96 well microtiter plates. A simplified scheme of this ELISA method is illustrated in Figure 2.

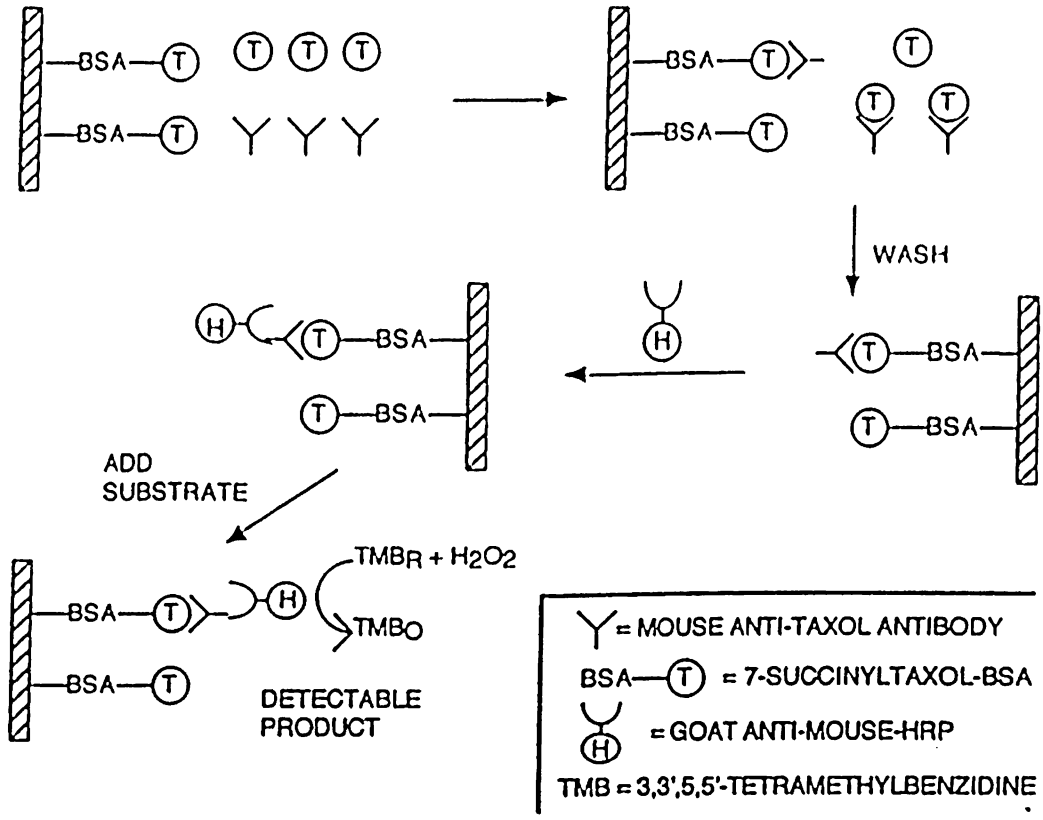


Figure 2. Flow chart of ELISA method

Microtiter high binding plate (Corning 25801) wells were coated for 2 h at 37°C with 100 µl per well of BSA-7-SucTax conjugate (5 µg/ml) in coating buffer. The plates were then washed four times and blocked for 1 h at 37°C with blocking buffer, then washed and air-dried. Physiological fluids (whole and parotid saliva, plasma, and plasma ultrafiltrate) were spiked with Tax-MeOH solution (0.5 mg/ml, 585 µM) and mixed with monoclonal Tax-Ab solution. The wells of low binding microtiter plates (Corning 25880) were used for mixing in order to obtain a serial dilution at a corresponding concentration range from 667 ng/ml (0.13 µl Tax-MeOH per well) to 1 pg/ml. After the uniform addition of monoclonal Tax-Ab at the optimal concentration of 78 ng/ml in PBS-T-20-BSA (0.2% BSA in PBS-T-20, pH 7.4; 0.01 M) in amounts of 100 µl per well, the plates were again incubated for 1 h at 37°C. High binding plates were washed four times with a wash buffer and air-dried. Then, mixed solutions from low binding plates were carefully transferred in the amount of 100 µl per each well so the antibody could freely compete with the binding between free and immobilized antigen an additional 1 h at 37°C. Plates were washed four times with wash buffer and horseradish peroxidase labeled Gt x Mo-HRP conjugate (1:15,000 dilution, 100 µl) in blocking buffer was added to each well and incubated at 37°C for 1 h. The plates were washed four times with nanopure water before adding 100 µl of microwell peroxidase substrate solution to each well for the colorimetric end point. 50 µl of 1 N HCl was added to each well after 10 min to stop the reaction and the plates were read in a 96 Kinetic Vmax Microplate reader at 450 and 650 nm.

Taxol Ab test for taxol analogues

All assays were performed under the same conditions for taxol detection described in the previous paragraph with the exception of a higher concentration of taxol analogues used prior to the serial dilution step, so that more accurate detection over the concentration range with the highest sensitivity could be achieved. Taxol analogues (dissolved in DMSO at a concentration of 2.5 mM) were adjusted to the final concentration of 50 μg/ml in blocking buffer in order to obtain a serial dilution in the concentration range from 50 μg/ml (about 30 μl of 2.5 mM solution in 1.5 ml buffer) to 48 pg/ml. Subsequent steps exactly followed the ELISA procedure previously described.

Results

The typical titration curve for the calculation of apparent binding constants for taxol antibodies is shown in Figure 3, where the optical density is plotted vs. antibody concentration. The Ab concentration is converted from the units μg/ml to nM.

The binding constant was estimated by taking the reciprocal of the half the maximum response. Binding constant calculated for taxol antibody towards the antigen BSA-7-SucTax conjugate was in the order of 10^9 M^{-1} (3.3×10^9). The main experimental variables used for the optimization of the assay were the concentrations of Tax-Ab and BSA-7-SucTax conjugate in sodium bicarbonate buffer, and the dilution of Gt x Mo-HRP conjugate in PBS-T-20-BSA. The pH of the samples might also be considered as another parameter. The sensitivity of the ELISA slightly increased as the alkalinity of our samples decreased (from pH 8.2 to the final pH 7.0) but this effect also improved the linearity of the absorbance (and the inhibition) plot, where the absorbance was plotted against the measured range of the concentration of Tax-MeOH in the physiological fluids. Optimal conditions, which led to the lower limit of detection (LOD), were determined by optimizing of the previous factors. They are as follows:

- concentration of BSA-7-SucTax [5 μg/ml]
- concentration of Tax-Ab [78 ng/ml]
- [1:15,000] dilution of Gt x Mo-HRP enzyme conjugate in PBS-T-20-BSA.

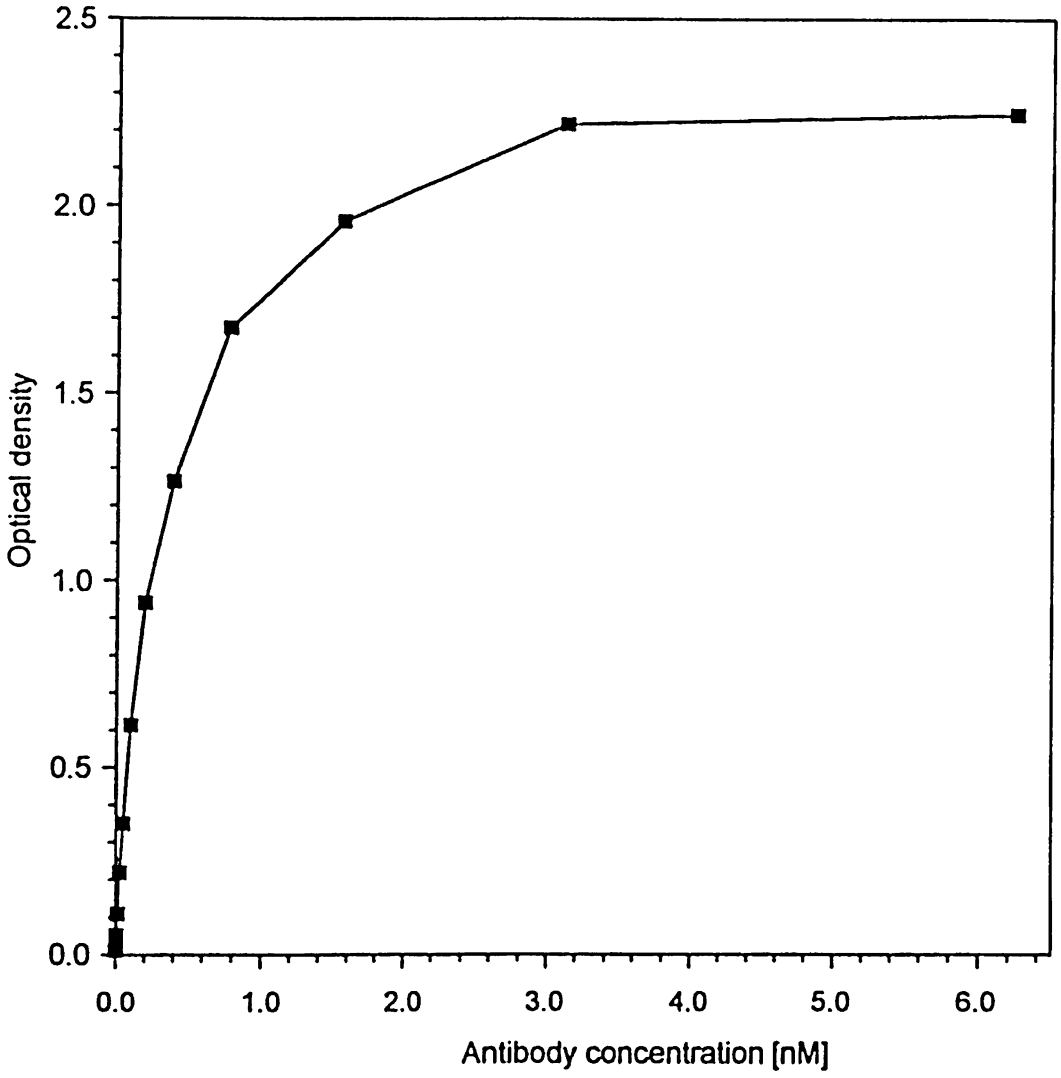


Figure 3. Titration curve for taxol antibody to BSA-7-SucTax conjugate (Tax-Ab at 10 μg/ml)

Figures 4 and 5 illustrate the optimization process for the experimental variable antigen and Tax-Ab.

Samples of taxol-spiked individual and pooled human physiological fluids (pH 7.0) were then analyzed under these optimal conditions. Our results for different physiological fluids in individual and pooled matrices are presented in Table 1.

Sample	Matrix	Detection
Saliva (whole)	Individual	Subnanogram (<50 pg/ml)
(whole)	Pooled	Subpicogram (<1 pg/ml)
(parotid)	Pooled	Subpicogram (<1 pg/ml)
Plasma ultrafiltrate	Individual	Subnanogram (<2 pg/ml)
	Pooled	Subnanogram (<3 pg/ml)
Plasma	Individual	Subnanogram (<350 pg/ml)
	Pooled	Subnanogram (<200 pg/ml)

Table 1 Limit of detection of taxol in human physiological fluids for different matrices

The results indicate detection of taxol at sub-nanogram concentration levels. The implications of these results are twofold. First, the limit of detection in plasma is much lower than previously published results by J. Leu with 40 ng/ml, or T. Willey, who reports 10 ng/ml. Second, multiple physiological fluids (plasma ultrafiltrate and salivary fluids) have been investigated. The range for quantitation is approximately five orders in magnitude (from pg/ml to 100 ng/ml).

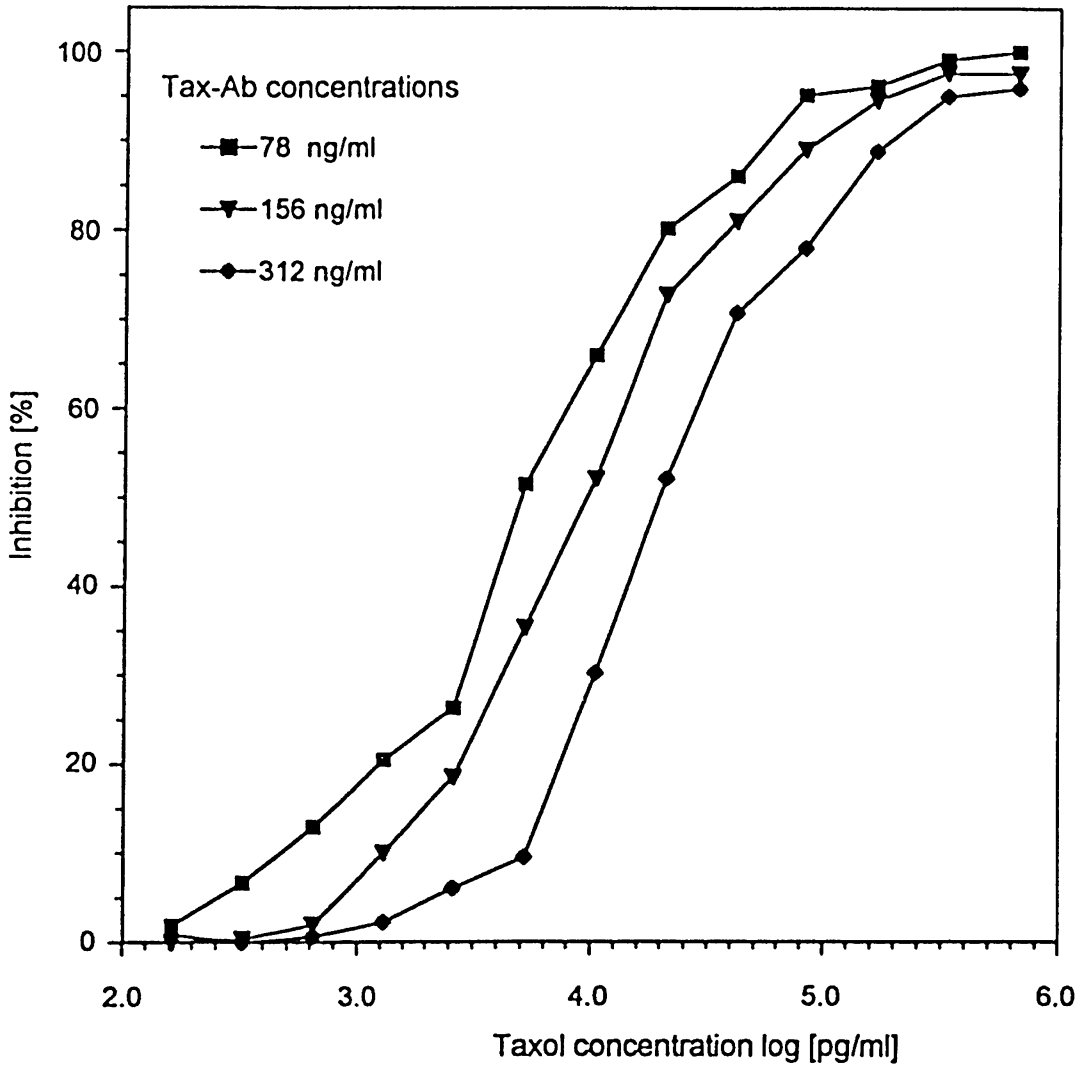


Figure 4. Optimization process for taxol antibody

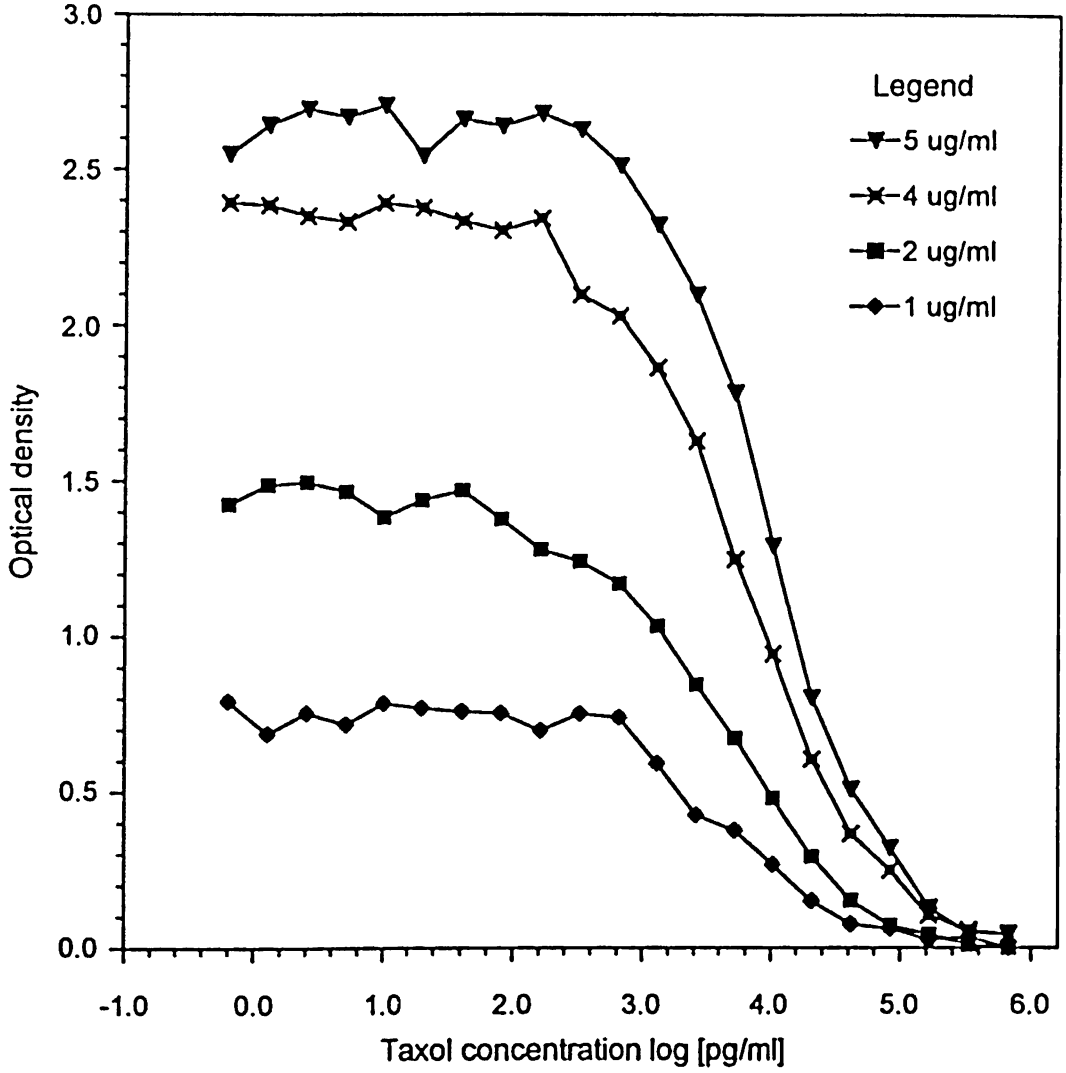


Figure 5. Optimization process for BSA-7-SucTax conjugate

Figure 6 represents the inhibition curve for taxol concentration in plasma ultrafiltrate at pH 7.0 on a logarithmic scale revealing a dynamic range of about five orders of magnitude with a sub-nanogram detection limit.

Since all these experiments were done in microtiter plates only in duplicate, there is no significance to the variance. The main source of the dispersion could be systematic and the sampling error.

Figure 7 shows the corresponding inhibition curve for taxol concentration in the salivary fluid at pH 7.0. Also in this medium, the broad dynamic range exceeds five orders of magnitude with a sub-picogram detection limit.

Figure 8 displays a calibration curve for paclitaxel in a plasma matrix with minimal sub-nanogram detection.

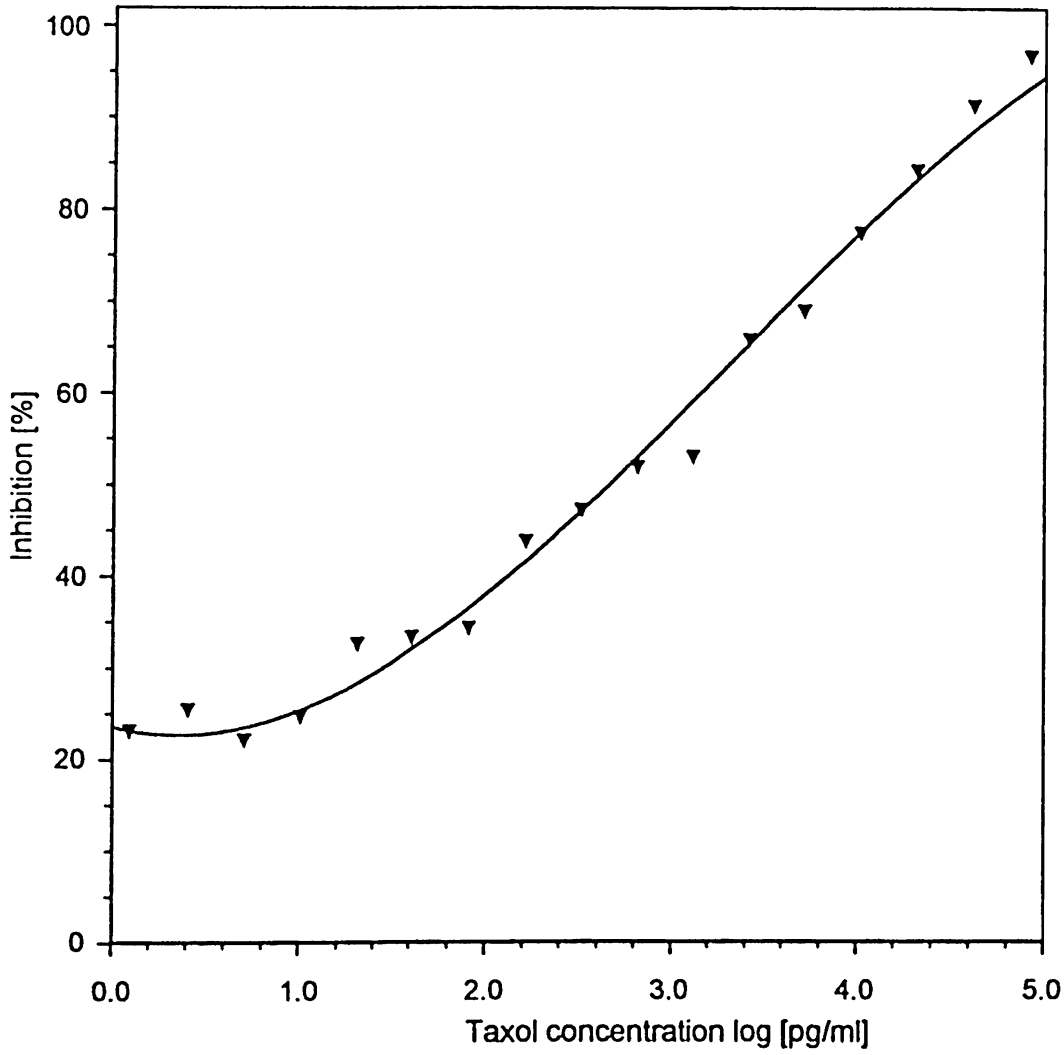


Figure 6. Taxol calibration curve in ultrafiltrate matrix (pH 7.0)

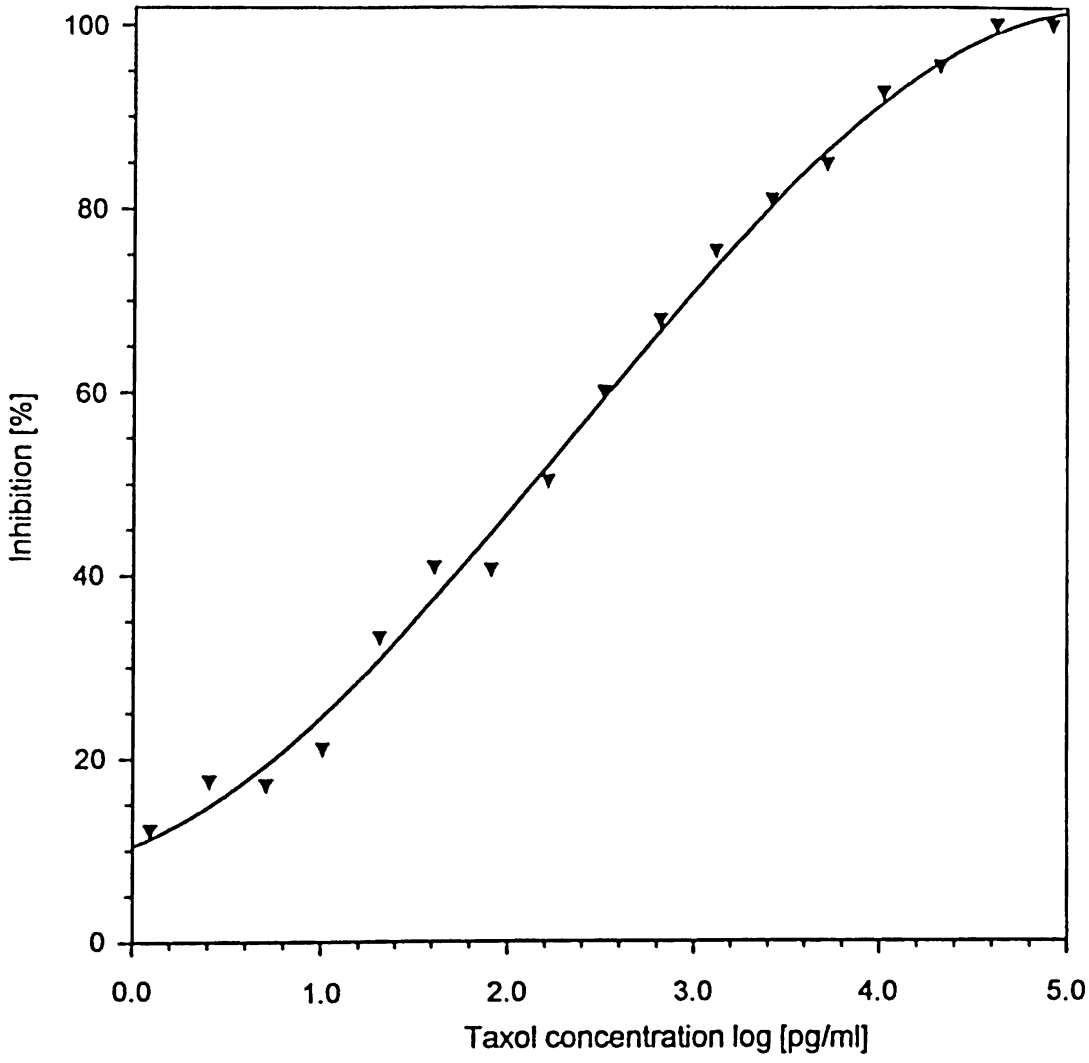


Figure 7. Taxol calibration curve in saliva matrix (pH 7.0)

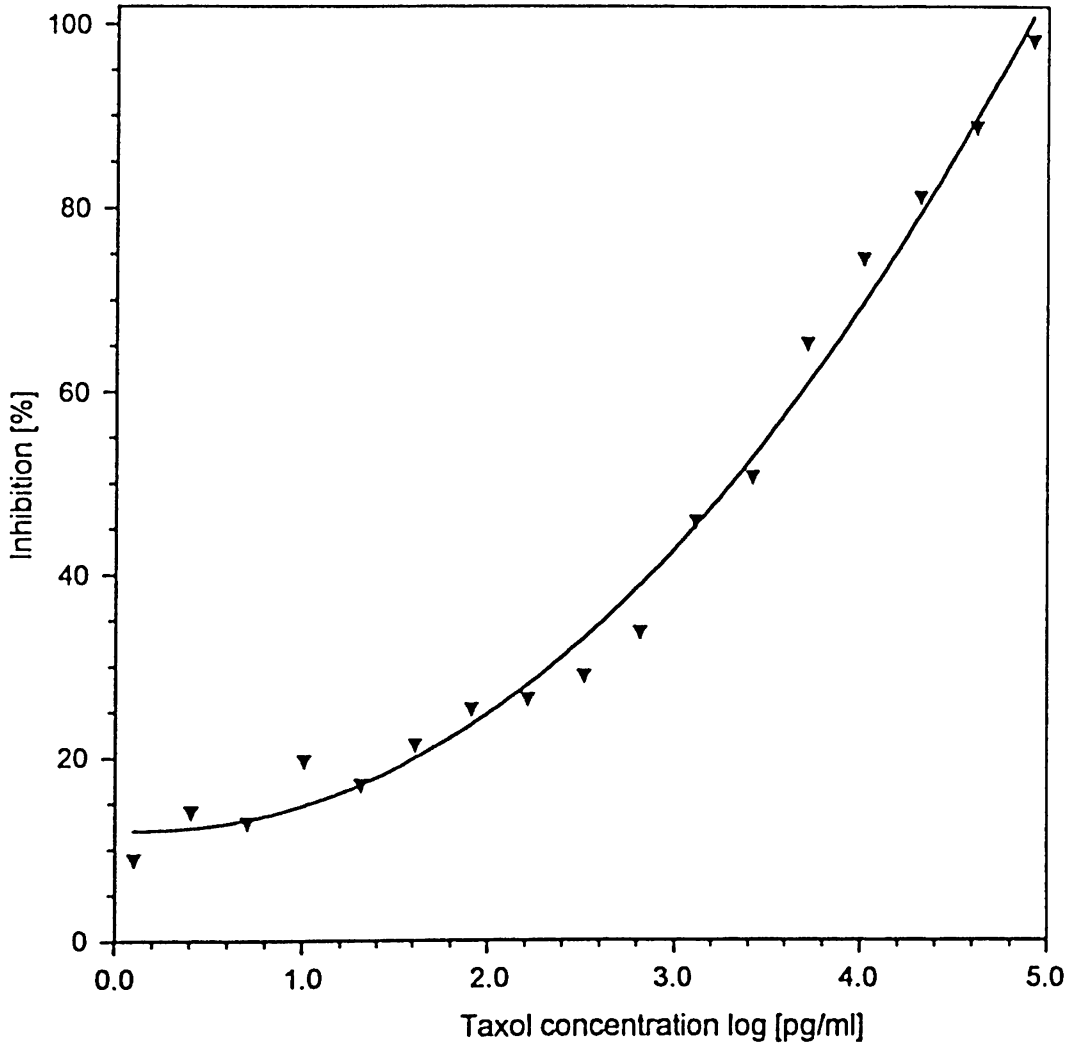


Figure 8. Taxol calibration curve in pooled plasma matrix (pH 7.0)

This assay was then applied to the taxol determination of blood and salivary samples obtained from a cancer patient treated with taxol that was administered at 250 mg/m^2 doses in 500 ml of 5% dextrose by a three-hour intravenous infusion. The body surface area (m^2) is calculated separately for adults and children by the sliding ruler provided by the Adria Pharmaceutical using weight (kg or lb) and height (cm or in). The samples were collected at predetermined time intervals during the I.V. infusion (0, 0.5, 1, 2, and 3 h) and after the I.V. infusion (0.5, 1, 2, and 12 h). Samples of saliva and plasma ultrafiltrate were diluted in a 1:100 ratio in PBS buffer solution. Plasma samples with an anticipated higher concentration of taxol were then diluted in 1:1000 ratio in the same solution. All samples were kept at temperature of -40°C.

Recovery of the ELISA method was performed with three measurements of known concentrations for each sample type. The samples were spiked with taxol to final concentrations of 0.2, 2.0, and 20 ng/ml and applied to the first twelve wells of the microtiter plate. Recovery results (% mean and RSD) are summarized in Table 2.

Concentration (ng/ml)	**Saliva** % (mean RSD)	**Plasma** % (mean RSD)	**Ultrafiltrate** % (mean RSD)
0.2	99.0 13.6	117 31.6	97.5 11.3
2.0	100 38.8	97.5 4.1	106 17.5
20	97.0 11.3	96.5 15.5	107 9.8

Table 2 Taxol recovery from spiked samples of physiological fluids
(RSD = Relative standard deviation = 100 (standard deviation) / mean)

Clinical applications of the ELISA method for detection of taxol in human biological fluids are illustrated in the next four figures, together with predicted values obtained by computer simulation at the 98% protein binding level.

Multiple sample analyses were performed repeatedly so the deviations could be calculated and plotted together with the mean values. We have to keep in our mind the quality of the human physiological fluids. The spiked samples were relatively 'clean', i.e. no other drugs were administered. When the 'real' samples were taken, it was an entirely different situation, because the patient had already taken several different drugs and their interference with taxol could be significant. The trends of the plotted curves are basically in reasonable agreement with the predicted, simulated values. As expected, the standard deviations are higher, when detection moves to a lower concentration level. The prime source of variation could be a sampling error with the consequential step of serial dilution and systematic error.

Figure 9 presents the summary of taxol detection by ELISA method in various biological fluids. Plasma samples support the anticipated higher concentration of taxol on sub-μg/ml level, while the saliva and ultrafiltrate samples are detected in the ng/ml concentration range. Taxol concentrations are then plotted against the adequate nonlinear pharmacokinetic model (Walle *et al.*, 1995; Gianni *et al.*, 1995; Wu *et al.*, 1996) to compare a differences between measured and simulated values.

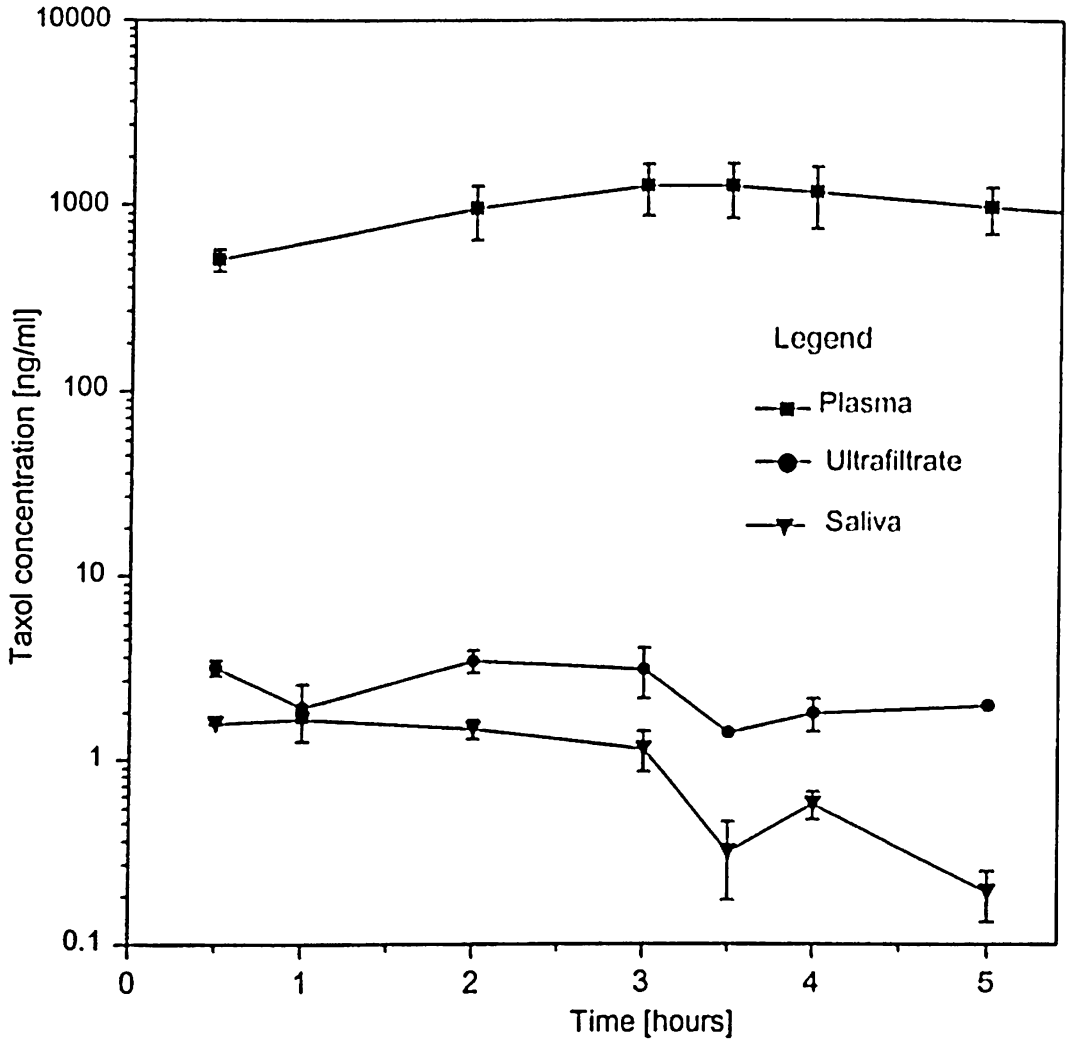


Figure 9. Detection of taxol in human biological fluids by ELISA method

The curves illustrated in Figure 10 (salivary fluids), Figure 11 (ultrafiltrate), and Figure 12 (plasma) are in a mutual agreement with simulated model and strongly support the feasibility of this assay for the further clinical pharmacokinetic studies including a larger number of patients treated with this drug.

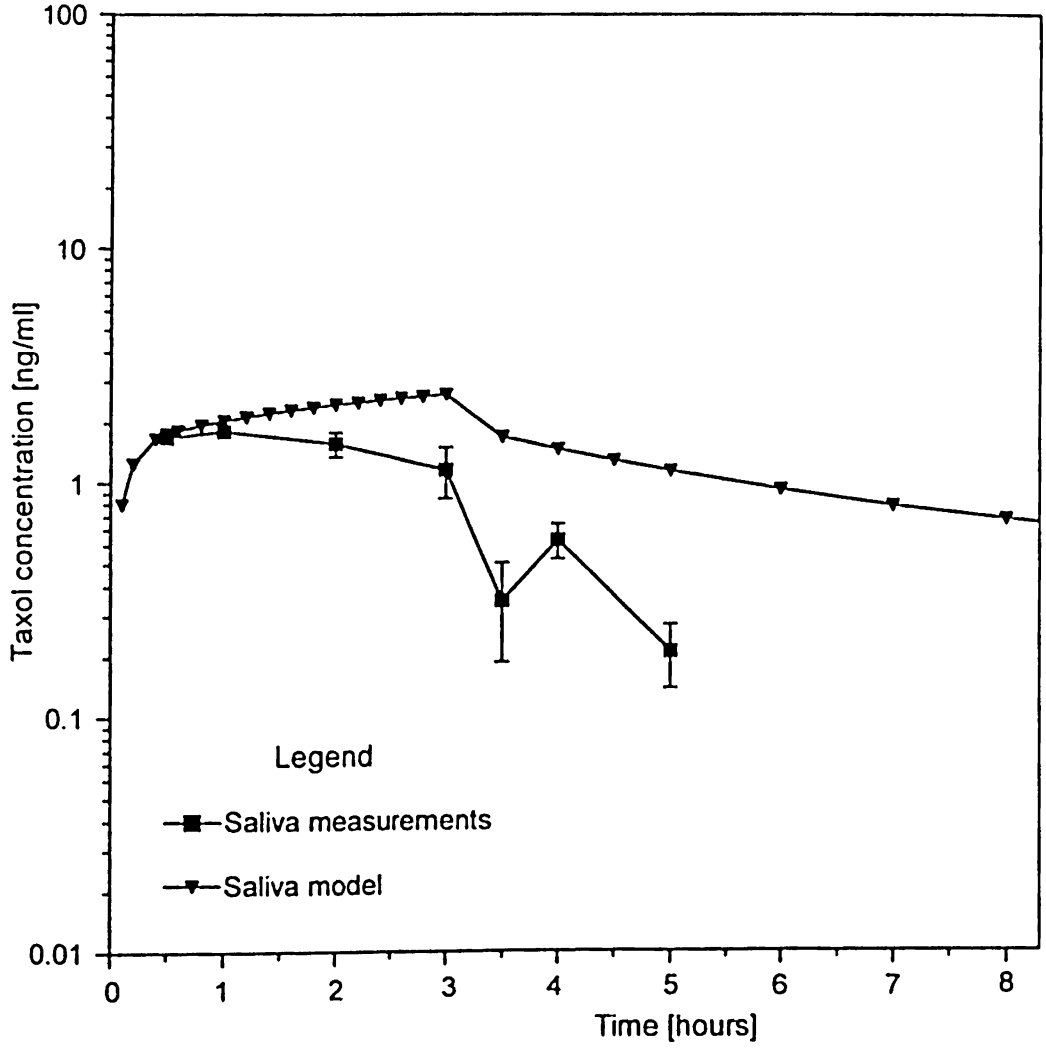


Figure 10. Simulated model with the actual detection of taxol in salivary fluids

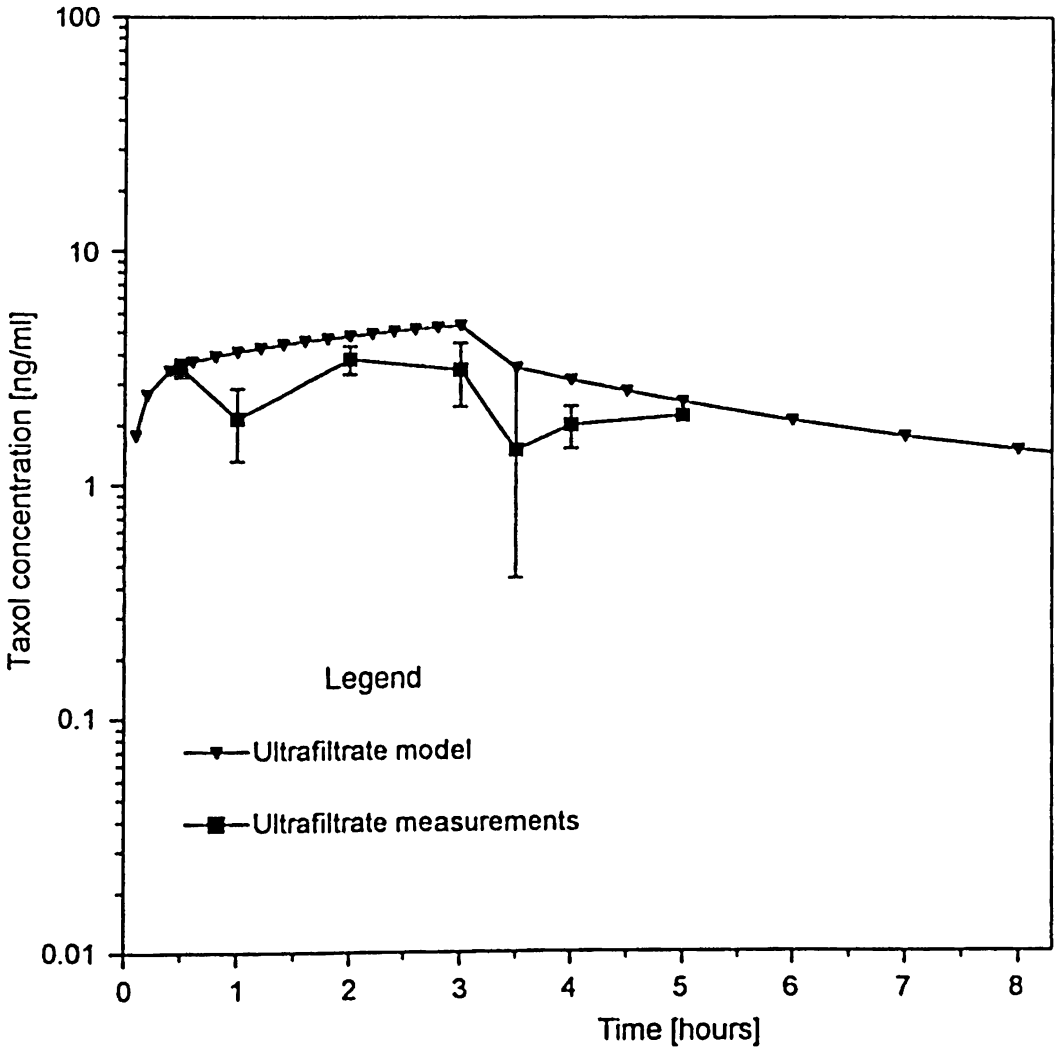


Figure 11. Simulated model with the actual detection of taxol in ultrafiltrate

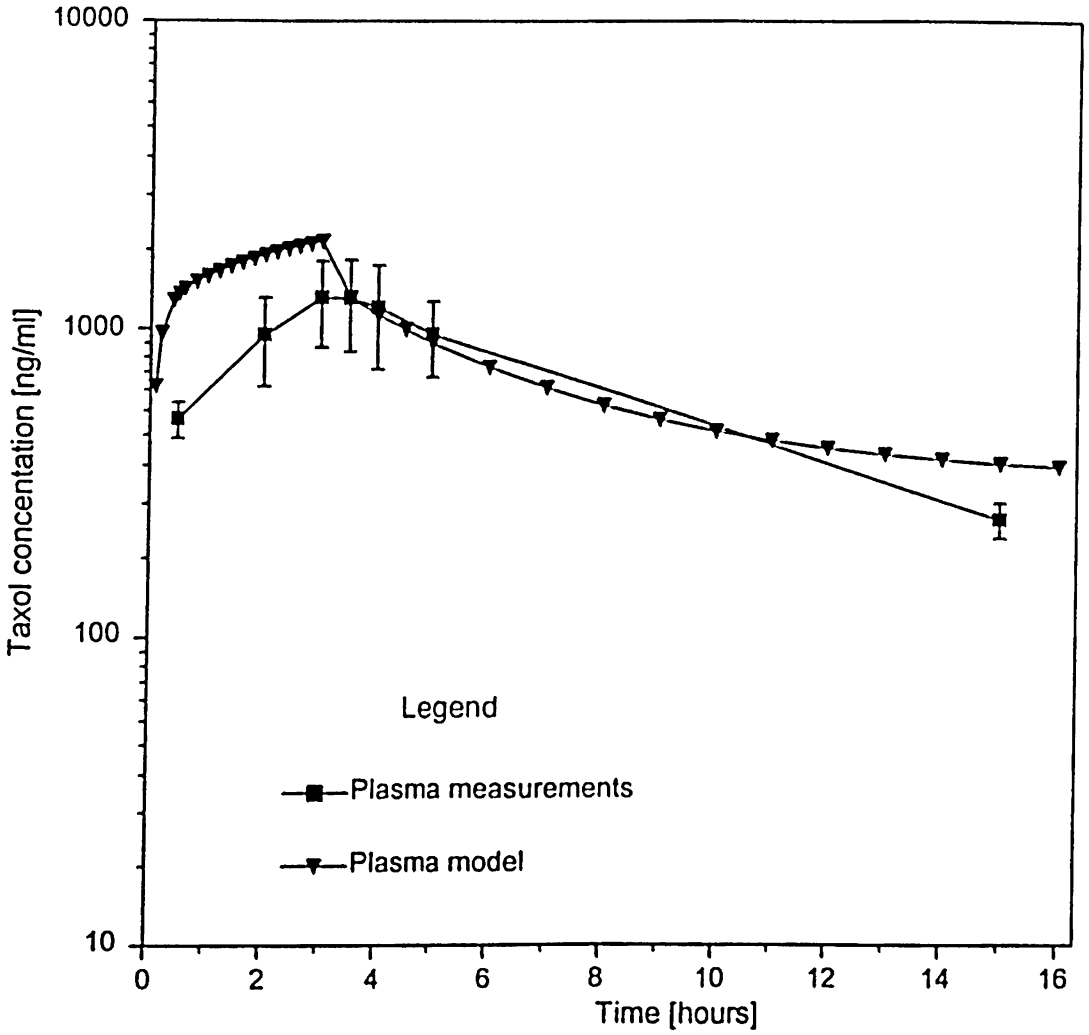


Figure 12. Simulated model with the actual detection of taxol in plasma

When the same ELISA analysis was performed on nine taxol analogues, quite strong interactions between monoclonal taxol antibody and three analogues were observed. Structures of these analogues (YBL1226-1, YBL1109-67, and YBL1024) are shown in Figure 13, while the inhibition curves for these three analogues are illustrated in Figure 14. The detection limit is in the nanogram and subnanogram/ml level, while the slopes of the graphs represent the sensitivity. These analogues must form very feasible epitopes similar to the taxol parent compound for such a strong interaction (binding) with taxol antibody.

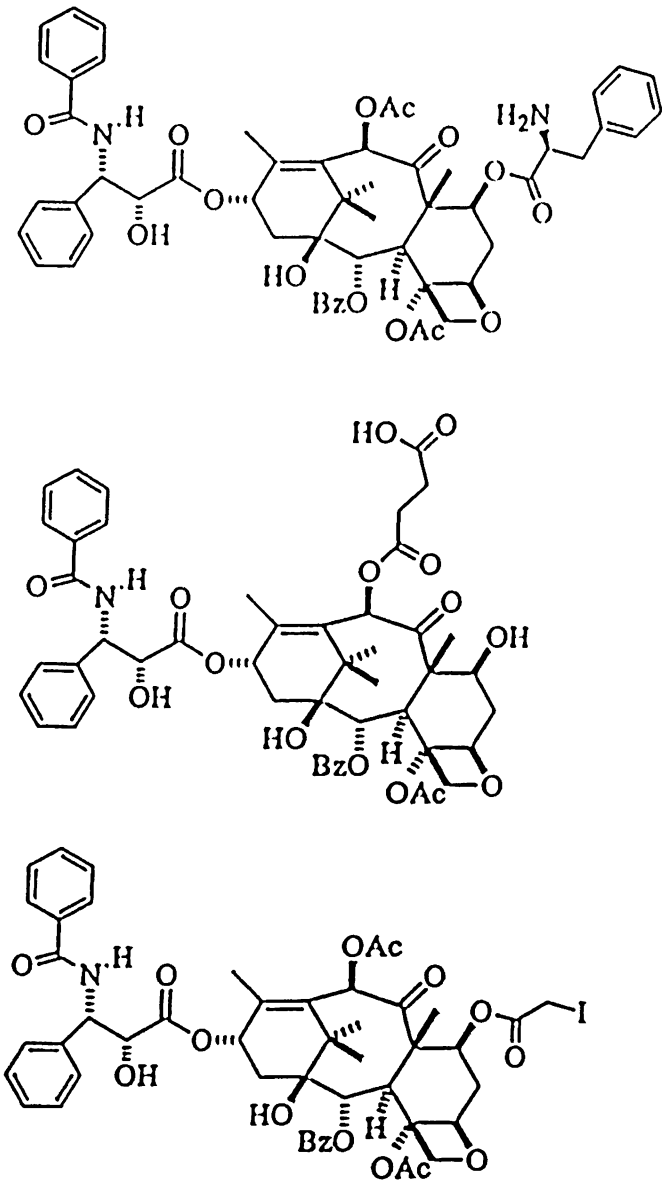


Figure 13. Structures of taxol analogues with high response towards Tax-Ab

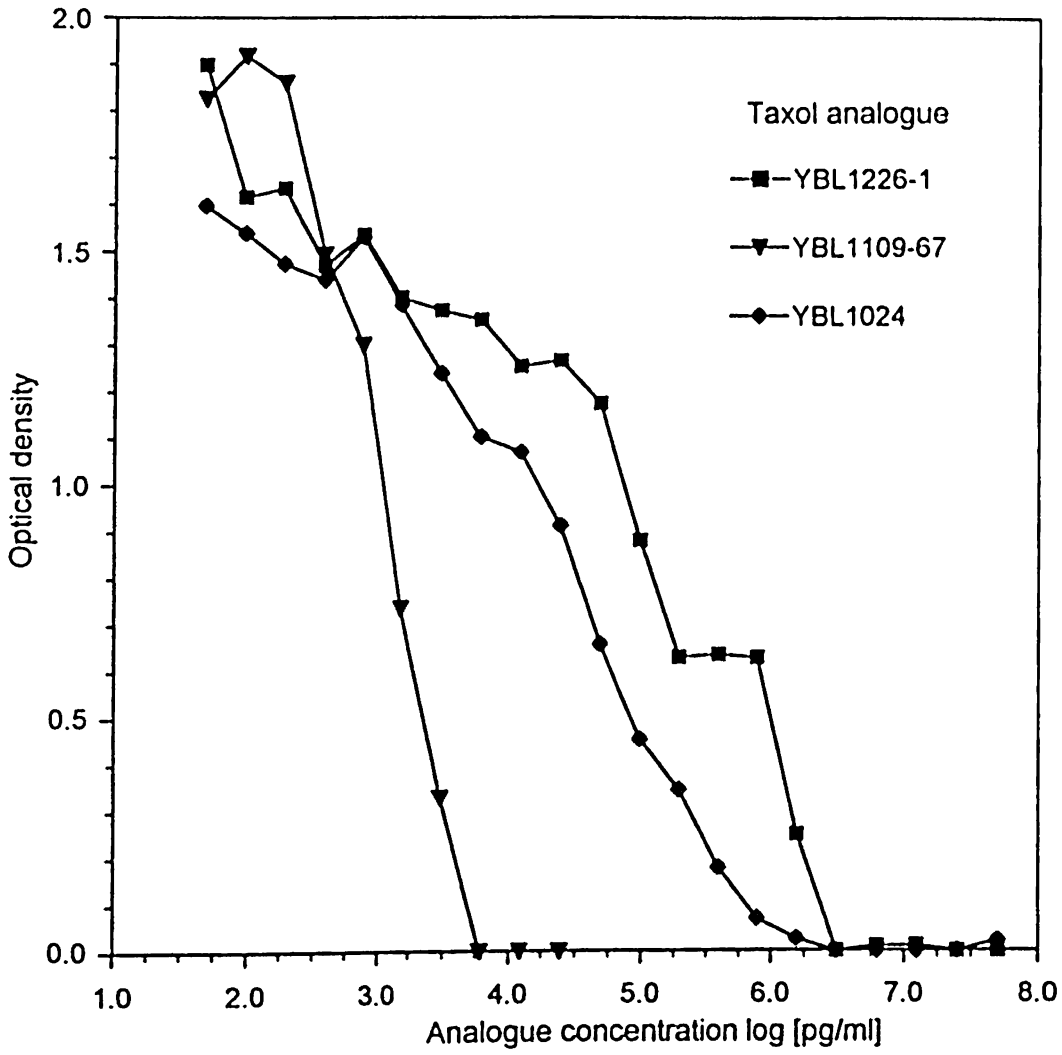


Figure 14. Calibration curves for the active taxol analogues

It is also important to consider the possible interference of taxol metabolites on the assay. Biliary excretion and hepatic metabolism have been described as the main elimination routes for taxol clearance. *In vitro* studies with human liver microsomes and tissue slices show that taxol is metabolized primarily to 6α-hydroxytaxol (Egorin, 1995), inactive against several cell lines, by the cytochrome P450 isozyme and to two minor metabolites, 3'-p-hydroxytaxol and 6α, 3'-p-dihydroxytaxol by another P450 isozyme. The concentration vs. time curve of the principal taxol metabolite in plasma follows the same general pattern as taxol (Wu *et al.*, 1996), but metabolite concentrations are far below the corresponding concentrations of the parent drug (less than 10%). Since the formation of this major metabolite (6α-hydroxytaxol) accounts for only a fraction of the total taxol concentration in human plasma, the effect of the metabolite on the assay results can be considered minimal even if metabolite binding is strong.

Discussion

Taxol is the most active anticancer drug introduced in cancer chemotherapy in the last ten years. Although the pharmacokinetic-pharmacodynamic relationship of this drug has been firmly established (Walle *et al.*, 1995; Wu *et al.*, 1996), the free, protein unbound and biologically active taxol could not be used for such studies due to the lack of highly specific analytical methodology. Since taxol is highly bound to human serum proteins [93%-98%] and protein binding may be variable among cancer patients, it is anticipated that pharmacokinetic studies directly evaluating the free, bioactive taxol species could further increase the therapeutic index of this promising drug (Gianni, *et al.*, 1995). This study presents the evaluation of the free, bioactive taxol species in a cancer patient using the ELISA technique. Development of an ELISA method with multivariate optimization to achieve such a low detection limit was essential.

Present HPLC methods available for the analysis of taxol concentrations in the physiological fluids have detection limits far above the levels required for the accurate determination of the taxol free, protein unbound forms. S. Auriola reports the detection limit of 50 ng/ml, while T. Willey and A. Sparreboom reached 10 ng/ml.

The very low detection limit and a broad range of sensitivity of the ELISA procedure not only enables the determination of the free biologically active taxol species, but would also allow pharmacokinetic studies of taxol administered at low dose

continuous infusion (Carbone *et al.*, 1995), which until now could not be performed using current HPLC methods. The feasibility of the ELISA method was clinically tested on samples from a cancer patient treated with this drug.

Although the clinical results are limited, they indicate the potential of this method for future clinical pharmacokinetic studies using biologically active taxol species. The results suggest higher protein binding of taxol at lower plasma concentration which will require additional studies already planned in our laboratories. Results from application of ELISA method to taxol analogues lead us into more detailed further investigation of the analogues, using a broad database from National Cancer Institute (NCI), Bethesda, Maryland with additional data from Department of Medicinal Chemistry, The University of Kansas, Lawrence.

In summary, a highly sensitive and specific ELISA analytical method has been developed with the determination limits of sub-nanomole taxol concentrations in physiological fluids. The preliminary results indicate its feasibility for clinical pharmacokinetic studies using free, biologically active taxol species. This method, currently tested in clinical studies involving larger number of cancer patients treated with taxol, is aimed in the direction of further improvement of the therapeutic index of this drug.

References

Auriola, S.O.K.; Lepisto, A.M.; Naaranlahti, T.; Lapinjoki, S.P. Determination of Taxol by high-performance liquid chromatography-thermospray mass spectrometry, *J. Chromatogr.*, **1992**, *594*, 153.

Brzezny, A.L.; Kalous, A.; Hajek, R.; Slavik, M. Clinical activity of Taxol in the treatment of non-small cell lung cancer (NSCLC). Book of abstracts. Annual Cancer Research and Training Symposium, Lawrence, KS, 10:1, **1997.**

Carbone, D.P.; Rosenthal, D. Continuous infusion paclitaxel with radiation therapy for locally advanced solid tumors. Book of abstracts. Paclitaxel: From Nature to Clinic, Riverside, CA, **1995**, 107.

Cardellina, J.H. HPLC separation of taxol and cephalomannine. *J. Liq. Chromatogr.*, **1991**, *14*, 659.

Chan, K.C.; Alvarado, A.B.; McGuire, M.T.; Muschik, G.M.; Issaq, H.J.; Snader, K.M. High-performance liquid chromatography in micellar electrokinetic chromatography of taxol and related taxanes from bark and needle extracts of *Taxus* species. *J. Chromatogr.*, **1994**, *657*, 301-306.

Egorin, M.J. Pharmacokinetics and pharmacodynamics of Paclitaxel. Book of abstracts. Paclitaxel: From Nature to Clinic, Riverside, CA, 1995, 1-14.

Gianni, L.; Kearns, C.M.; Giani, A.; Capri, G.; Vigano, L.; Locatelli, A.; Bonadonna, G.; Egorin, M.J. Nonlinear pharmacokinetics and metabolism of paclitaxel and its pharmacokinetic/pharmacodynamic relationships in humans. *J. Clin. Oncol.*, **1995**, *13(1)*,180-190.

Gregory, R.E.; DeLisa, A.L. Paclitaxel: a new antineoplastic agent for refractory ovarian cancer. *Clin. Pharm.*, **1993**, *12(6)*, 401-415.

Grothaus, P.G.; Raybould, T.J.G.; Bignami, G.S.; Lazo, C.B.; Byrnes, J.B. An enzyme immunoassay for the determination of taxol and taxanes in *Taxus* sp. tissues and human plasma. *J. Immunol. Methods*, **1993**, *158(1)*, 5-15.

Grothaus, P.G.; Bignami, G.S.; O'Malley, S.; Harada, K.E.; Byrnes, J.B.; Waller, D.F.; Raybould, T.J.; McGuire, M.T.; Alvarado, A.B. Taxane-specific monoclonal antibodies: measurement of taxol, baccatin III, and "total taxanes" in *Taxus brevifolia* extracts by enzyme immunoassay. *J. Nat. Prod.*, **1995**, *58(7)*, 1003-1014.

Hajek, R.; Vorlicek, J.; Slavik, M. Paclitaxel (Taxol): a review of its antitumor activity in clinical studies. *Neoplasma*, **1996**, *43(3)*, 141-154.

Hamel, E.; Lin, C.M.; Johns, D.G. Tubulin - dependent biochemical assay for the antineoplastic agent taxol and application to measurement of the drug in serum. *Cancer Treat. Rep.*, **1982**, *66*, 1381.

Harvey, S.D.; Campbell, J.A.; Kelsey, R.G.; Vance, N.C. Separation of taxol from related taxanes in *Taxus brevifolia* extracts by isocratic elution reversed-phase microcolumn high-performance liquid chromatography. *J. Chromatogr.*, **1991**, *587*, 300-303.

Hempel, G.; Lehmkuhl, D.; Krumpelmann, S.; Blaschke, G.; Boos, J. Determination of paclitaxel in biological fluids by micellar electrokinetic chromatography. *J. Chromatogr. A.*, **1996**, *745(1-2)*, 173-179.

Jaziri, M.; Diallo, B.M.; Vanhaelen, M.H. Vanhaelen-Fastre, R.J.; Zhiri, A.; Beeu, A.G.; Homes, J. ELISA for the detection and the semi-quantitative determination of taxane diterpenoids related to taxol in Taxus sp. and tissue cultures. *J. Pharm. Belg.*, **1991**, *46(2)*, 93-99.

Kalous, A.; Brzezny, A.L.; Hajek, R.; Slavik, M. Clinical antitumor activity of Taxol in the treatment of advanced breast cancer. Book of abstracts. Annual Cancer Research and Training Symposium, Lawrence, KS, 10:6, 1997.

Kingston, D.G.; Chaudhary, A.G.; Chordia, M.D.; Gharpure, M.; Gunatilaka, A.A.L.; Higgs, P.I.; Rimoldi, J.M.; Samala, L.; Jagtap, P.G. Synthesis and biological evaluation of 2-acyl analogues of paclitaxel (taxol), *J. Med. Chem.*, **1998**, *41*, 3715-3726.

Koeller, J.M.; Dorr, R.T. Pharmaceutical issues of Paclitaxel. Supplement to *The Annals of Pharmacotherapy*, **1994**, *28(5)*, 5.

Leu, J.G.; Jech, K.S.; Wheeler, N.C.; Chen, B.X.; Erlanger, B.F. Immunoassay of taxol and taxol-like compounds in plant extracts. *Life Sci.*, **1993**, *53(12)*, 183-187.

Li, C.; Yu, D-F; Newman, R.A.; Millas, L; Hunter, N.R; Wallace, S. Development of highly efficacious water-soluble polymer-taxol conjugate. *Proc. Annu. Meet. Am. Assoc. Cancer Res.*, **1997**, 1731:A

Mathew A.E.; Mejillano, M.R.; Nath, J.P.; Himes, R.H.; Stella, V.J. Synthesis and evaluation of some water-soluble prodrugs and derivatives of taxol with antitumor activity. *J. Med. Chem.*, **1992**, *35*, 145-151.

Natale, R. Combination Carboplatin & Paclitaxel in non-small cell lung cancer. Book of abstracts. Paclitaxel: From Nature to Clinic, Riverside, CA, 1995, 33-33.

Ozols, R. Carboplatin plus Paclitaxel in advanced ovarian cancer. Book of abstracts. Paclitaxel: From Nature to Clinic, Riverside, CA, **1995**, 58-60.

Rizzo, J.; Riley, C.; von Hoff, D.; Kuhn, J.; Phillips, J.; Brown, T. Analysis of anticancer drugs in biological fluids; determination of taxol with application to clinical pharmacokinetics. *J. Pharm. Biomed. Anal.*, **1990**, *8(2)*, 159-164.

Rowinsky, E.K.; Onetto, N.; Canetta, R.M.; Arbuck, S.G. Taxol: The first of the taxanes, an important new class of antitumor agents. *Semin. Oncol.*, **1992**, *19*, 646-662.

Schiff, P.B.; Fant, J.; Horwitz, S.B. Promotion of microtubule assembly *in vitro* by taxol. *Nature*, **1979**, *277*, 665-667.

Sparreboom, A.; De-Bruin, P.; Nooter, K.; Loos, W.J.; Stoter, G.; Verweij, J. Determination of paclitaxel in human plasma using single solvent extraction prior to isocratic reversed-phase high-performance liquid chromatography with ultraviolet detection. *J. Chromatogr.*, **1998**, *705(1)*, 159-164.

Stasko, M.W.; Witherup, K.M.; Ghiorzi, T.J.; McCloud, T.G.; Look, S.; Muschik, G.M.; Issaq, H.J. Multimodal thin layer chromatographic separation of taxol and related compounds from *Taxus brevifolia*. *J. Liq. Chromatogr.*, **1989**, *12*, 2133-2138.

Walle, T; Walle, U.K.; Kumar, G.I.; Bhalla, K.N. Taxol metabolism and disposition in cancer patients. *Drug Metab. Dispos.*, **1995**, *23(4)*, 506-512.

Wani, M.C.; Taylor, H.L.; Wall, M.E.; Coggon, P.; McPhail, A.T. Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.*, **1971**, *93*, 2325-2327.

Willey, T.A.; Bekos, E.J.; Gaver, R.C.; Duncan, G.F.; Tay, L.K.; Beijnen, J.H.; Farmen, R.H. High-performance liquid chromatographic procedure for the quantitative determination of Paclitaxel (Taxol) in human plasma. *J. Chromatogr.*, **1993**, *621(2)*, 231-238.

Wu, J.; Stobaugh, J.; Slavik, M. Clinical pharmacokinetic studies of taxol. Book of abstracts. Annual Research Forum, Wichita, KS, 6:8, **1996**.

Chapter 2

MEASUREMENT OF THE BIOLOGICAL ACTIVITY OF PROTEINS IMMOBILIZED ON SENSORS

Abstract

This study addresses a new class of labelless immunosensors with direct response to the binding of a ligand to a receptor immobilized on a sensor electrode. The sensing mechanism is based on antigen-antibody direct binding reaction on an ultrathin (~ 20 Å) discontinuous platinum film. The detection at sub-ng/ml levels of *Staphylococcal Enterotoxin B* (SEB) has been accomplished with anti-SEB antibody immobilized on the sensor surface. The success of the sensor in responding to antigen-antibody direct binding depends on the quality of the ultrathin Pt film structure and chemical immobilization of the antibody. In this project of anti-SEB antibody immobilization on the sensor surface, we used radioactive ^{125}I labeling as the detection system of active antibody. The results show that the final concentration of bioactive immobilized antibody is approximately 1 pmol/cm^2. This technology has the potential for a generic low cost diagnostic assay and could be used in the medical and food industry.

Introduction

The main objective of this study is to demonstrate the sensing mechanism of a new ultrathin platinum film immunosensor based on antigen-antibody direct binding reaction. *Staphylococcal enterotoxin B* (SEB) globular protein is used as the ligand, while the anti-SEB antibody (SEB-Ab) is a receptor for this model system with radiometric detection.

The sensing mechanism based on antigen-antibody direct binding reaction on the surface of the sensor could be potentially used as the baseline for the conductimetric sensors that can detect directly the antigen-antibody binding events without the need for any kind of labeling (enzyme-, radio-, and so on) together with the multiple wash steps essential in most immunoassays. A change of conductivity with a proper baseline can measure and quantify a direct transduction of the ligand-receptor binding process. But the serious disadvantage is that the measured impedance change is very small, so the signal has to be greatly amplified. The impedance decreases with the formation of antigen-antibody complex. If the complex is built by small concentration changes (addition of the same amount of antigen to immobilized bioactive antibody), the step concentration change could be detected as the comparable step in the impedance change.

Very unique is the metal electrode, which comprises a discontinuous ultrathin platinum film with a thickness of approximately 20 to 25 Å. The film forms a porous layer containing separated platinum islands. The dimensions of the metal islands, as

well as the spaces between them are comparable (the same dimensional range) to antibodies (SEB-Ab) immobilized on the electrode surface surrounded by platinum islands. In this way a mosaic network of metal islands and bound protein is formed. The metal layer is attached to an insulating substrate of silicon dioxide coated on silicon wafer. SEB-Ab is covalently attached to the silicon dioxide surface. The whole sensor is a square with a side length of 1 cm. Figure 1 shows the cross-section of the ultrathin platinum film sensor.

The conductance is expected to be the most sensitive within roughly 100 Å of the surface (Kasapbasioglu *et al.*, 1993). The success of the sensor in responding to antigen-antibody direct binding depends on the quality of the deposited ultrathin platinum film structure and chemical immobilization of the antibody. The film morphology can be controlled by deposition conditions.

The immobilization of the antibodies has been already described by many authors (Kasapbasioglu *et al.*, 1993; DeSilva *et al.*, 1995; Johnsson *et al.*, 1995; Le *et al.*, 1995) with minor differences. The procedure generally results in about 15% to 20% of the attached antibodies likely to be biologically active. The goals of this project are to investigate and optimize the strategy of the antibody immobilization chemistry and the sensitive radiometric detection system to approximate the amount of biologically active antibodies on the surface of the sensor.

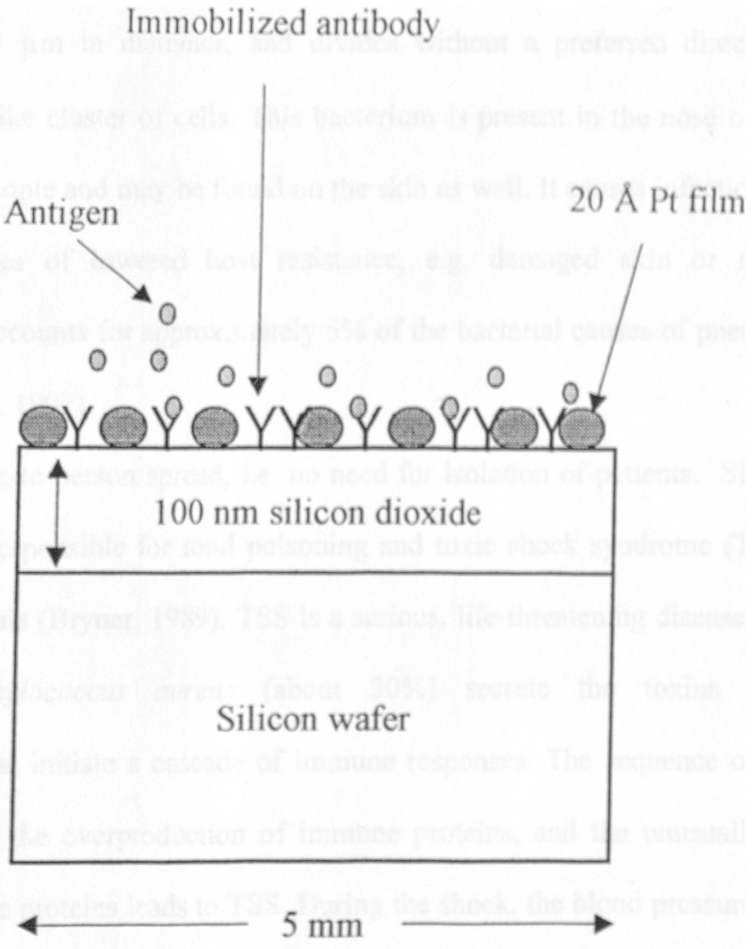


Figure 1. Cross-section profile of the immunosensor

Staphylococcal enterotoxin B (SEB)

Staphylococcal enterotoxin B (SEB) is isolated from the cell membrane of the bacterium *Staphylococcal aureus*. The staphylococcal cell is typically spherical (coccus), about 1 μm in diameter, and divides without a preferred direction to produce a grape-like cluster of cells. This bacterium is present in the nose of about 30% of healthy people and may be found on the skin as well. It causes infection most commonly at sites of lowered host resistance, e.g. damaged skin or mucous membranes and accounts for approximately 5% of the bacterial causes of pneumonia (Humphreys *et al.*, 1997).

There is no person-to-person spread, i.e. no need for isolation of patients. SEB is a globular protein responsible for food poisoning and toxic shock syndrome (TSS) in humans and animals (Bryner, 1989). TSS is a serious, life-threatening disease. Some strains of *Staphylococcus aureus* (about 30%) secrete the toxins (called superantigents) that initiate a cascade of immune responses. The sequence of these reactions leads to the overproduction of immune proteins, and the unusually high production of these proteins leads to TSS. During the shock, the blood pressure drops dramatically to very low level, which can be fatal without fast antibiotic treatment.

Purified toxin is a white powder-like material that is readily soluble in water and salt solutions. The recombinant form of SEB is commercially available. It contains four high affinity-binding sites capable of interacting with the antibodies of several species. The protein molecule consists of a single polypeptide chain (about 300 amino

acids, 28 lysines) with the molecular weight about 35,000 Da. SEB is not destroyed by boiling (i.e. relatively heat stable) and is resistant to denaturing agents such as 4 M urea, 70% ethanol, and 0.1 M HCl. These properties, together with its toxicity, make the SEB protein an excellent candidate as the antigen molecule in the antigen-antibody interaction.

Experimental

Material

The sensors coated with the ultra-thin Pt film and the device similar to microtiter plate cylindrical well (with the diameter 0.5 cm) were custom made in Microfabrication Application Laboratory, The University of Illinois at Chicago, IL 60607. Staphylococcal enterotoxin B (SEB), anti-SEB rabbit antibodies (SEB-Ab), Bovine Serum Albumin (BSA) and Tween 20 for blocking the solution were purchased from Sigma Chemical (St. Louis, MO). Purification of the antiserum sample was performed on chromatography columns (bed size 0.8 x 4 cm) obtained from Bio-Rad (Richmond, CA). Reactigel®-6X for affinity support, Immuno-Pure Gentle Ag/Ab elution Tris buffer, and bicinchoninic acid (BCA) protein assay reagents were purchased from Pierce (Rockford, IL). The substrate solution for Horseradish Peroxidase (HRP), Tetramethyl benzidine (TMB) was from Kirkegaard and Perry Inc. (Gaithersburg, MD). Goat anti-rabbit-HRP (Gt x Rb-HRP) was purchased from Organon Teknika-Cappel Corp. (Durham, NC). The microtiter plates came from Corning (Corning, NY). 3-Mercaptopropyltrimethoxysilane (MTS) was obtained from Petrarch Systems (Bristol, PA). The heterobifunctional cross-linker N-gamma-maleimidobutyryloxysuccinimide ester (GMBS) was purchased from Calbiochem (San Diego, CA). Centrifree columns for separation of free from protein bound microsolute (molecular cut off 10 - 30 kDa) were purchased from Amicon Inc. (Beverly, MA). Toluene (Aldrich, Milwaukee, WI) was kept over 3 Å molecular

sieves. Bolton-Hunter reagent (total activity 260 μCi) for labeling with ^{125}I was obtained from ICN Pharmaceuticals, Inc. (Irvine, CA). The radioactivity of each sensor was individually measured with a scintillation gamma counter, 1282 CompuGamma, CS, LKB Wallac (Finland). The nanopure water system was from Barnstead Nanopure II (Boston, MA). All other reagents used were of analytical grade.

Characterization of anti-SEB antiserum (SEB-Ab)

To one vial of lyophilized powder of the rabbit SEB-Ab antiserum was added 2 ml of deionized water to reconstitute the product to the initial concentration of 47 mg/ml.

The binding characteristics of the SEB antibody (SEB-Ab) towards its antigen SEB coated on the microtiter plates were studied. A standard procedure to obtain titration curve (optical density vs. antibody concentration) via direct, noncompetitive ELISA was performed on high-binding polystyrene ELISA 96 well microtiter plates.

The plate wells were coated for 2 h at 37°C with 100 μl per well of SEB (5 μg/ml) in sodium bicarbonate (coating) buffer (pH 9.5; 0.1 M). The plates were then washed four times with phosphate buffered saline (PBS; pH 7.4; 0.01 M) containing 0.5% (v/v) Tween 20 (PBS-T-20) and blocked for 1 h at 37°C with 100 μl of blocking buffer (PBS-T-20, containing 0.2% BSA of ELISA grade (98-99%)). After washing the plate four times with wash buffer (PBS-T-20) and drying them at room temperature, SEB-Ab at the concentration of 10 μg/ml was added to the first two wells and serially diluted by two wells over the microtiter plate. The last four wells

were used as blank to eliminate the background signal. Plates were incubated at 37°C for 1 hr and washed again with wash buffer (pH 7.4; 0.01 M). Then, Gt x Rb-HRP enzyme conjugate in blocking buffer (1:10,000 dilution, 100 µl) was added, the plates incubated for 1 hr at 37°C and washed four times with nanopure water before adding 100 µl of microwell peroxidase substrate (TMB) solution uniformly to each well. The colorimetric reaction was stopped after 15 min by adding 50 µl of 1 N HCl to each well, and the absorbance was measured on Kinetic Vmax Microplate Reader (Molecular Devices Corp., Menlo Park, CA) connected to a Windows based PC.

The avidity of the Rb-SEB-Ab antiserum was then established from the titration curve using the antibody concentration at the half of the optical density.

Purification of SEB-Ab antiserum

Further antiserum purification was performed in order to obtain more specific (but still polyclonal) antibodies towards the SEB antigen with expected higher apparent binding constant. SEB was adjusted to the working concentrations of 1 mg/ml and 1 mg then coupled to 1 ml of Reactigel®-6X solution (1,1'-carbonyldiimidazole activated crosslinked 6% agarose) according to the manufacturer's instructions. The sample was mixed for 30 hr at 4°C in coupling buffer (carbonate buffer, pH 8.5) by mechanical inversion. The imidazole liberated during protein coupling contributes to the absorbance at 280 nm. Therefore, the determination of protein coupling by measuring absorbance A_{280} is invalid.

The amount of the protein present in solution before and after the derivatization can be measured by bicinchoninic acid (BCA) protein assay, so the coupling efficiency E [%] can be calculated according to the next equation.

E [%] = 100 [(the mass of the protein bound to the column)/(the total protein mass)]

After the coupling reaction, the beads are transferred to the small chromatography column (bed size 0.8 x 4 cm). The gel is incubated with a blocking buffer (1 M Tris-HCl; pH 8.0) up to 4 hr to deactivate the active groups and then washed with coupling buffer to remove blocking buffer and to elute the protein molecules that were nonspecifically attached. Finally the coupled gel was washed with 0.01 M phosphate (PB) buffer (pH 7.4) and stored in the same buffer at 4°C. The polyclonal SEB antiserum sample was passed through the affinity column matrix containing immobilized SEB at very low flow rate of about 0.1 ml/min. The unbound and weakly bound fractions of SEB-Ab were eluted from the column by excessive washing with 0.01M Tris buffer (pH 7.4). The bound fraction of SEB-Ab was then eluted using high ionic strength Immuno-Pure Gentle Ag/Ab elution (Tris) buffer. These collected fractions were concentrated and dialyzed in 0.01M PBS (pH 7.4) containing 0.05% azide. Affinity purified antibodies were characterized against the SEB antigen and used in the radioiodination procedure. The ELISA method was similar to the previously described procedure.

Radioiodination procedure

Radiolabeling of affinity purified SEB-Ab by Bolton-Hunter reagent was a modification of the procedure described by Bolton and Hunter (1973). The authors developed a method for iodination in which the molecules are not themselves exposed to the ^{125}I (I^-) solution. This method does not require tyrosyl residues, but rather lysines that are more frequently distributed in peptide and protein molecules.

The Bolton-Hunter reagent is the iodinated N-hydroxysuccinimide (NHS) ester of 3-(4-hydroxyphenyl) propionic acid. The NHS-ester end of the reagent reacts with primary amines of the molecule to be iodinated (Rudinger *et al.*, 1973).

The Bolton-Hunter reagent provided 260 μCi in 13 μl volume (concentration 20.0 mCi/ml). The radioiodination was performed according to the recommended protocol in a specially designated hood. The separation of labeled proteins and radioactive volatile iodine was done using the activated charcoal filter (iodine trap) and desalting chromatography column packed with Sephadex G-10 and equilibrated with an 0.1% gelatin-buffer solution. This is very important procedure, since even a small amount of free radioactive iodine remaining with the radiolabeled protein leads to a large error in the evaluation of adsorbed protein. Reconcentration of the radiolabeled SEB-Ab was completed using the Amicon centrifuge filters (molecular cut off = 10 kDa).

Antibody immobilization on a sensor

Covalent immobilization techniques most often require prior 'pre-treatment' (activation, hydrolysis) of the support material as the initial step. The procedure is a modification of the method described by DeSilva *et al.*, 1995.

The sensor with the Pt film surface was immersed in concentrated sulfuric acid for 15 min, followed by rinsing several times with nanopure water, and boiled for an additional 15 min in nanopure water. Finally, the substrates were dried at 70°C.

Treating the surface with a vapor of 2.0% v/v solution of MTS in dry toluene for 30 min in a small vial and under the inert (nitrogen) atmosphere carried out the silanization process of the film surface. Nitrogen was blown into the vial for an additional 2 min, then the vial was sealed and left in the oven at 50°C for 30 min. The sensors were annealed in oven at 70°C for an additional 90 min.

Next, the surface was immersed for a 15 min in 10 mM mercaptoethanol/ethanol solution, rinsed in dry toluene and allow to air dry. The silanized sensor surface was then treated for 1 hr with a heterobifunctional cross-linker, GMBS (2 mM in absolute ethanol) solution and washed with PBS (pH 7.4; 0.1 M) buffer.

Activation of the sensor surface prior to immobilization of purified or radioactive antibody was performed on the set of 12 sensors. Six of them were used for the immobilization of radiolabeled antibody and the other half was used for the attachment of affinity purified SEB-Ab and as blank (i.e. without any antibody) to

eliminate the background signal. The identical conditions, during the activation process for all sensors, ensure the minimum variation (experimental error) between the sensors.

A volume of 50 µl of affinity purified or radiolabeled SEB-Ab (standard concentration of 100 µg/ml) in PBS solution was delivered to the activated substrate and was allowed to incubate overnight at 4°C.

Then the film was rinsed with an excess of PBS buffer to elute the unbound SEB-Ab portion and was left for 1 hr in 10% monoethanolamine in PBS solution to deactivate the surface. In the case of ^{125}I-labeled antibodies, the amount of surface-bound protein was determined using a scintillation gamma counter.

The scheme of the immobilization procedure for covalent attachment of proteins to a silica surface is illustrated in Figure 2.

The sensor was then washed with the 5.0% (w/w) solution of sodium dodecyl sulfate (SDS; $[CH_3–(CH_2)_{10}–CH_2–O–SO_3^-]\,Na^+$) in PBS (pH 7.4; 0.1 M) in separate culture dishes and the counts per minute (cpm) read again using the gamma counter. This procedure confirms the strength of the immobilized SEB-Ab attachment, i. e. covalent bond (from the linker to antibody) or some kind of non-specific attachment (adsorption).

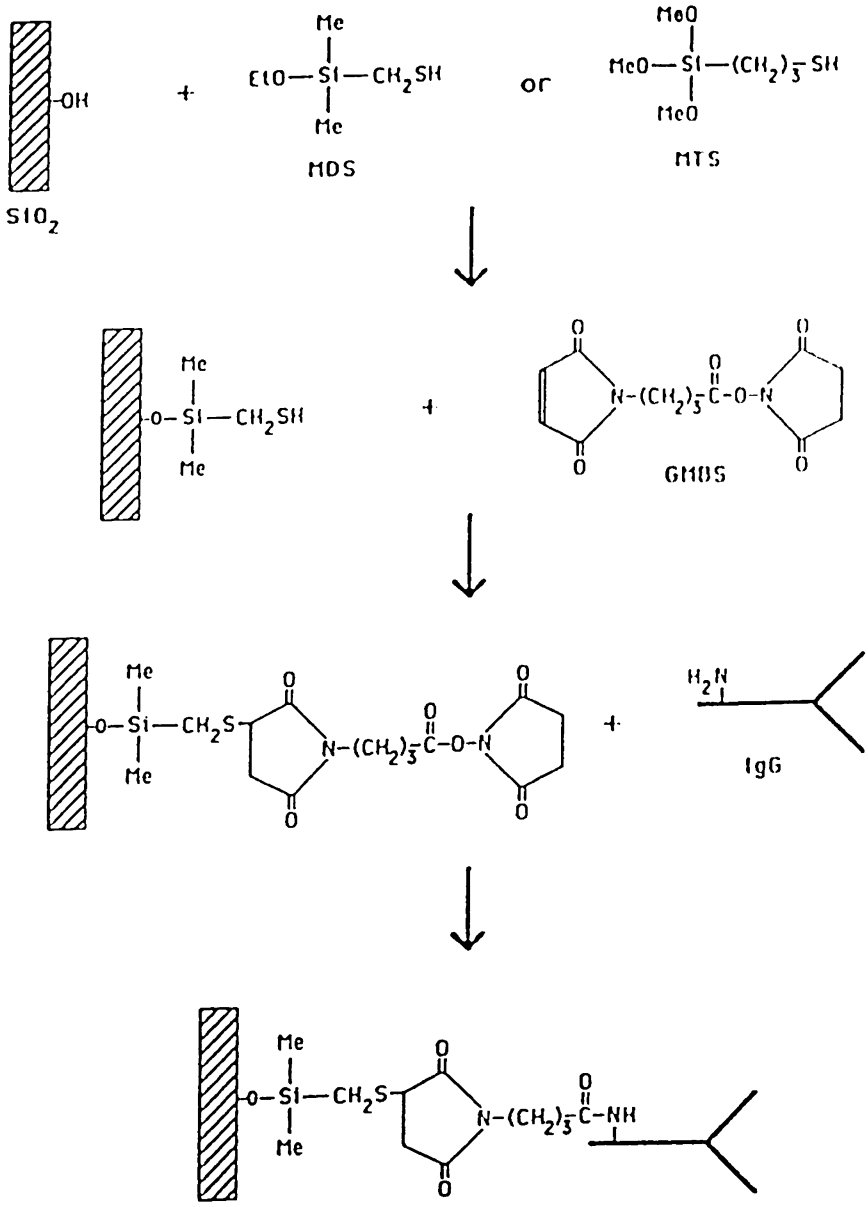


Figure 2. Immobilization procedure for covalent attachment of SEB-Ab to a silica surface

Results and discussion

The calculated apparent binding constants based on the ELISA methods are summarized in Table 1.

SEB-Ab antiserum	K_a [$10^8 M^{-1}$]
Initial	3.8
Affinity purified	6.1
Radiolabeled	0.8

Table 1 Apparent binding constants of SEB-Ab antiserum to SEB

Affinity purification of the initial SEB-Ab antiserum increases the pool of mono-specific (polyclonal) antibodies with the expected elevated apparent binding constant. Radiolabeling process in any form (Bolton-Hunter procedure, Iodobeads reagent and direct iodination) has the opposite effect on the protein due to the changes in its conformation. The comparison (with the corresponding non-labeled protein) shows that even with a 'gentle' Bolton-Hunter reagent, the biological activity of radiolabeled antibody decreased approximately 8 times.

Figure 3 shows the titration curve for the initial SEB-Ab antiserum to SEB antigen (concentration of 5 μg/ml).

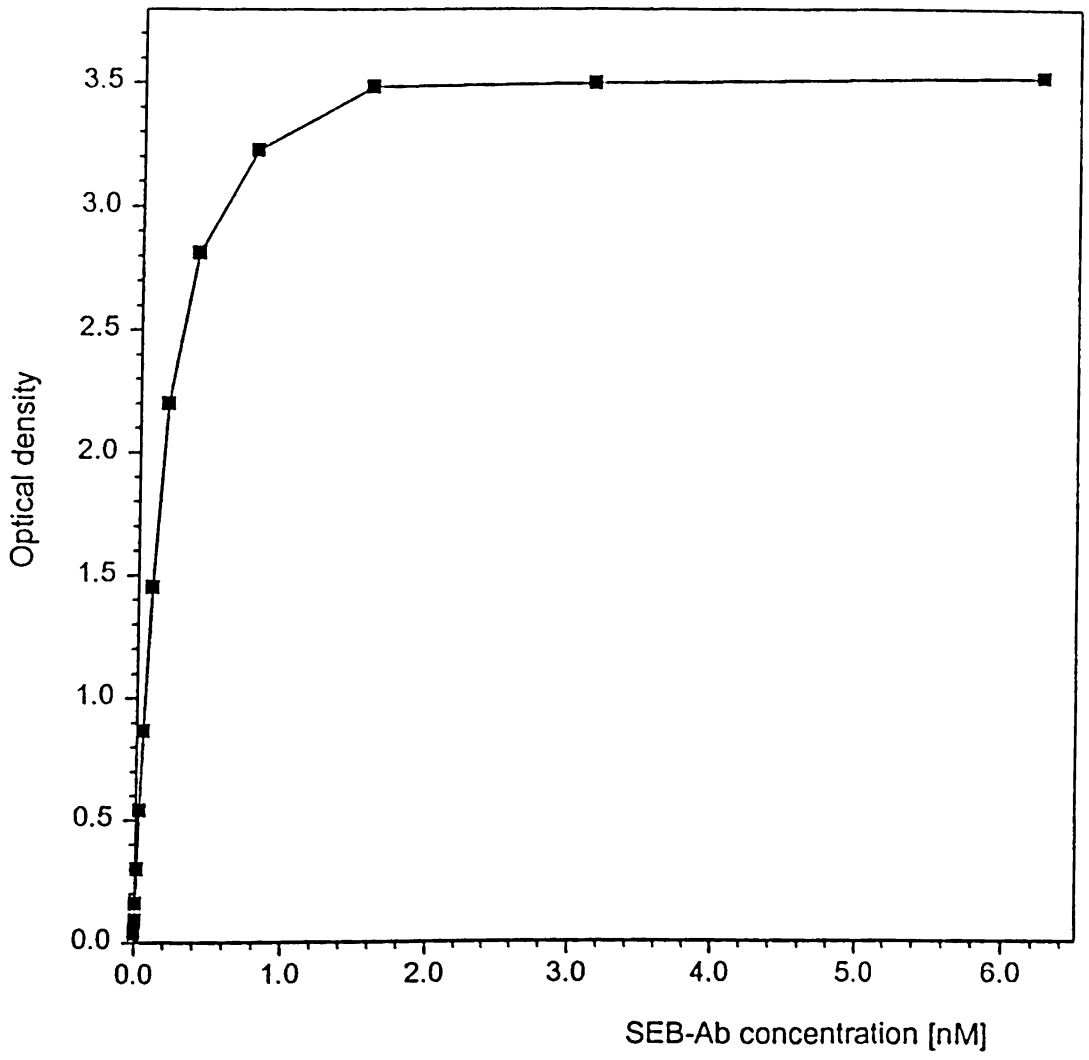


Figure 3. Titration curve for SEB-Ab antiserum to SEB (conc. 5 μg/ml)

The coupling efficiency of the affinity chromatography column was about 94.2%. Affinity purification leads into higher mono-specificity of SEB-Ab but yields only about 3% of the initial amount of protein introduced on the chromatography column. Starting with 2.0 ml of the SEB-Ab antiserum (47 mg/ml), the affinity purification produced 3.140 mg of the purified protein in the concentration of 1.12 mg/ml, i.e. a total of 2.803 ml. This is the expected result.

The radiolabeling procedure with the Bolton-Hunter reagent (followed by the purification and reconcentration process using Amicon centrifuge filters) led to the total amount of 1.147 mg of radiolabeled SEB-Ab (starting with 2.0 ml of 1.0 mg/ml affinity purified antibodies) in 1.623 ml. The final concentration was found to be 0.707 mg/ml using the bicinchoninic acid (BCA) protein assay. Efficiency of the radiolabeling process provided 1,864,157 counts per minute (cpm) for the total amount of labeled antibody, which revealed that one antibody molecule out of approximately 2.3×10^9 were actually radiolabeled. It was a sufficient level because a relatively large surface area (0.2 cm^2) was actually coated.

The amount of 50 µl of the SEB-Ab (100 µg/ml) was transferred directly on the surface of the sensor, i.e. 5.0 µg, which is approximately 31.3 pmol of SEB-Ab.

Table 2 shows the radioactivity of radiolabeled antibody (100 µg/ml), which was measured directly on the gamma counter.

Sensor	cpm
Treated surface + Ab	120
Treated surface + Ab	124
Treated surface + Ab	128
Non-treated surface + Ab	75
Non-treated surface, no Ab	68
Treated surface, no Ab	72

Table 2 Radioactivity measurements

The background is about 72 cpm and the average signal, produced by radiolabeled immobilized antibodies, is approximately 52, with the precision ± 3.27 cpm (standard deviation). Since 5 μg/50 μl of radioactive antibody added to the surface equals 31.3 pmol counted for 8,125 cpm, 52 cpm corresponds to 0.196 pmol per the coated surface area (0.2 cm^2).

The experiment was then repeated to assure the reproducibility (precision) of the immobilization procedure. The second result showed almost identical amount of immobilized SEB-Ab, 1.035 $pmol/cm^2$.

The treatment with SDS solution has no significant effect (decrease) on the counts (verified by Student's statistical hypothesis test), which confirms the strong covalent attachment of the antibody to the crosslinker molecule.

The proper activation of the sensor surface (hydrolysis), silanization chemistry and the chemistry of immobilization of the antibody on the sensor exterior are sufficient procedures for the covalent attachment of proteins to a silica surface via the crosslinker. The area of the Pt film feasible for the immobilization of the SEB-Ab was about 0.2 cm^2 (circle with the 0.5 cm in the diameter). Since the average cpm (with the standard deviation) was 52 ± 3.27 cpm, the adequate amount of immobilized antibodies is approximately 0.20 pmol/0.2 cm^2, which means about 1 pmol/cm^2. Binding of the labeled antibody with ^{125}I was designated as the total binding.

Conclusion

The initial, purified and radiolabeled SEB-Ab Rb antiserum show different apparent binding constants, which is in good agreement with other authors (Iandolo *et al.*, 1988; Marrack *et al.*, 1990).

The affinity purification produced mono-specific antibody with expected higher apparent binding constant and yielded in about 3% of initial amount of the SEB-Ab.

The detection limit of the SEB-Ab covalently immobilized on the sensor surface to the antigen (SEB) is approximately 1.0 ng/ml. The success of the sensor in responding to antigen-antibody binding strongly depends on the quality of the ultrathin film and the immobilization method of the antibody. In this study of the radiolabeled SEB-Ab immobilized on the sensor surface, the results show approximately 1 pmol/cm^2 as the final concentration of bioactive immobilized antibody. The detection limit is lower than other published data.

Harteveld *et al.* described a sensor device in a flow injection system, where the SEB-Ab concentration, incubation times and flow rate were optimized. The detection limit for such a system was reported as 100 ng/ml. Other authors published even higher limits. Mukhin *et al.* reported a detection limit close to 5 µg/ml, while Strachan *et al.* described the limit of detection at 10 µg/ml.

This technology has a great potential for a low cost diagnostic assay and could be successfully used in the medical or food industry.

Milner *et al.* illustrated a practical application of similar sensor for detection of clinical mastitis by changes in electrical conductivity of foremilk before visible changes in milk. Changes in electrical conductivity of foremilk indicated the establishment of bacteria. When clots first appeared in foremilk, 90% cases of elevated bacteria occurrence were detected. All subclinical infections from *Staphylococcal aureus* were detected as well.

Other applications could include a rapid and accurate detection of staphylococcal infection in humans.

References

Bolton, A.E.; Hunter, W.M. The labelling of proteins to high specific radioactivities by conjugation to a ^{125}I-containing acylating agent – application to the radioimmunoassay. *Biochem. J.*, **1973**, *133*, 529-538.

Bryner, C.L. Recurrent toxic shock syndrome. *American Family Physician*, **1989**, *39*, 157-164.

DeSilva, S.M.; Zhang, Y.; Hesketh, P.J.; Maclay, G.J.; Gendel, S.M.; Stetter, J.R. Impedance based sensing of the specific binding reaction between *Staphylococcus enterotoxin B* and its antibody on an ultra-thin platinum film. *Biosens. Bioelectron.*, **1995**, *10(8)*, 675-682.

Harteveld, J.L.N.; Nieuwenhuizen, M. S.; Wils, E.R. Detection *of Staphylococcal Enterotoxin B* emloying a piezoelectric crystal immunosensor. *Biosens. Bioelectron.*, **1997**, *12*, 661-667.

Humphreys, H.; Keane, C.T. Methicillin-resistant *Staphylococcus Aureus* (MRSA) and vancomycin-resistant enterococci. *Lancet*, **1997**, *350*, 737-738.

Iandolo, J.J.; Tweten, R.K. Purification of Staphylococcal enterotoxins. *Met. Enzym.*, **1988,** *165,* 43-53.

Johnsson, B.; Lofas, S.; Lindquist, G.; Edstrom, A.; Hillgren R.M.; Hansson, A. Comparison of Methods for immobilization to carboxymethyl dextran sensor surfaces by analysis of the specific activity of monoclonal antibodies. *J. Mol. Rec.*, **1995,** *3,* 125-131.

Kasapbasioglu, B.; Hesketh, P.J.; Hanly, W.C.; Maclay, J.C. Esfahani, R.N. An impedance based ultra-thin platinum island film glucose sensor. *Sensors and Actuators,* **1993,** *14,* 749-751.

Le, D.; He, F.; Jiang, T.J.; Nie, L.H.; Yao, S.Z. A goat-anti-human IgG modified piezoimmunosensor for *Staphylococcus aureus* detection. *J. Microbiol. Meth.*, **1995,** *23,* 229-234.

Marrack, P.; Kappler, J. The Staphylococcal enterotoxins and their relatives. *Science,* **1990,** *248,* 705-711.

Milner, P.; Page, K.L.; Walton, A.W.; Hillerton, J.E. Detection of clinical mastitis by changes in electrical conductivity of foremilk before visible changes in milk. *J. Diary. Sci.*, **1996,** *79(1),* 83-86.

Mukhin, D.N.; Chatterjee, S. A receptor-based immunoassay to detect Staphylococcus enterotoxin B in biological fluids. *Anal. Biochem.*, **1997,** *245(2),* 231-217.

Rudinger, J.; Ruegg, U. Preparation of N-succinimidyl 3-(4-hydroxyphenyl) propionate. *Biochem. J.,* **1973,** *133,* 538-539.

Strachan, N.J.; John, P.G.; Millar, I.G. Application of rapid automated immunosensor for the detection of Staphylococcus aureus enterotoxin B in cream. *Int. J. Food. Microbiol.,* **1997,** *35(3),* 293-297.

Chapter 3

DESIGN OF NEURAL NETWORK MODELS FOR SCREENING ANTICANCER ACTIVITIES IN TAXOL ANALOGUES

Abstract

This study describes the procedure of a back-propagation neural network (BPNN) design for 50 related pharmaceutical compounds with 27 numerically quantified physical and chemical properties (feature vector). There are 40 compounds with known output in the training set. Based on the training data set and BPNN architecture, we can make meaningful and very accurate predictions of the anticancer activity for the 10 tested analogs. The selection of the system design depends greatly on the nature of the non-linearity to be modeled. For data sets containing periodicities and some pattern recognition, the results indicate that the BPNN is, in this case, more flexible and gives much better performance in comparison with the statistical discriminant analysis (Bayes' rule) based on the assumption of normally distributed inputs.

In this investigation, BPNN results in accuracy of 92% in predicting anticancer activity against ovarian and lung cancer, 94% for breast cancer and 90% for the index GI_{50}, while discriminant analysis results in accuracy of 58% for ovarian, 68% for lung, 76% for breast cancer, and 60% for the index GI_{50}. Correlation between predicted and actual outputs (activities) is often used as an additional parameter of model accuracy. In terms of correlation coefficient, BPNN accuracy is 0.831 for ovarian, 0.945 for lung, and 0.913 for breast cancer, while the index GI_{50} reveals an accuracy of 0.886.

Introduction

Over the last ten years, activity in the field of artificial neural networks (NN) has increased exponentially. This effort has led to an enormous volume of research publications, textbooks, and even creation of at least four new journals as outlets for work specifically related to this field. Data analysis is nothing new: it has been performed for many years, mostly by statistical methods. However, knowledge acquisition by the human brain is not performed by any statistical method. It has been clear for a long time that the human brain analyzes data and information quite differently from those statistical methods and that it processes a flood of data and becomes more efficient through learning.

The mathematical models and algorithms, that have been designed to imitate the information processing and knowledge acquisition methods of the human brain, are called neural networks. Artificial neural networks are computer models derived from a very simplified concept of the brain.

There is no universally accepted definition of a NN. Most people in this field would probably agree that a NN is a network of many simple processors (neurons, nodes), each having a small amount of local memory. Communication channels (connections) which usually carry numeric data, encoded by various means, connect the units. The units operate only on their local data and on the inputs they receive via the connection. The networks have a very unique set of characteristics. They are not programmed but trained by being repeatedly shown large numbers of examples for

the problem under consideration. Once trained, the concurrent architecture can provide good results in a relatively short time. Collection of data, pre-processing of the data, data analysis, and the systematic approach to optimize the NN model are all essential parts of NN design. The fundamental assumption of neural network theory is that the transfer signals are not linearly dependent on the net input. The problems handled by NN can be quite different. On the most general level they can be divided into four different categories.

Association: The system is able to reconstruct the correct pattern even if the input pattern is incomplete or somehow corrupted by a certain level of noise.

Classification: The goal is to assign all objects under investigation to appropriate classes (clusters) of objects, based on one or more properties that actually characterize a given class.

Mapping: This is a procedure of transformation of a multivariate space into another space of lower dimensionality.

Modeling: The search for the analytical function or a procedure (model) that will give a specified *n*-variable output for any *m*-variable input.

These procedures are mainly used for deciding whether a given multivariate input signal (called the feature vector) belongs to a certain class or category. Such a multidimensional vector can represent an audio signal, an optical image, a spectrum of any kind, a many-component chemical and physical analysis and so on.

The key attributes of neural network could be summarized as follows:

Learning from experience

The NN models are particularly suited to problems whose solution is very complex and difficult to specify, but providing an excess of data from which the response output could be learned.

Generalizing from examples

This important property of any self-learning system is the ability to extrapolate and interpolate for defined intervals from previous learning experience. Through a systematic approach, a NN model can be trained to give the correct response to data not previously encountered. This is often described as the ability to generalize on test data (Hopfield, 1982; Kohonen, 1988; Pao, 1989; Ripley, 1993).

Error tolerant system

This is an important property of NN to withstand errors at the input of the network, as well as errors in the weights. As error-tolerant, the network output will not deviate significantly from the target output based on the input pattern even if the pattern contains some errors, corrupted data or noise up to some limit.

Developing solutions faster

The NN model is trained by example. As long as examples are available and an appropriate design adopted, effective solutions can be reached far more quickly than through traditional approaches.

Computational efficiency

Training a NN model in the concurrent architecture is computationally intensive, but the memory requirement (of a fully trained model) for data testing is very modest.

Non-linearity

NN models can be trained to generate non-linear mapping with no assumption of the underlying data distribution and this often gives them an advantage in dealing with real complex problems. The NN will transform m-variable (multivariable) input into *n*-variable output (multiresponse). The input or output variables can be:

Real numbers - preferably scaled into <0,1> analog interval

Binary numbers, i.e. 0 or 1

Bipolar numbers, i.e. –1 or +1

The number of input and output variables is limited only by the available hardware, computation times, memory requirements, and type of NN design. The number of output variables is usually smaller than on the input side. The basic entity of any NN model is the formal neuron (Figure 1). Its action consists in summing all weighted inputs and transforming them into output signals via activation function. Although the activation (transfer) function can have several forms (Wasserman, 1989; Zupan *et al.*, 1993), the most commonly used is the sigmoid function which is characterized by the equation:

$$f(x) = y = 1/[1 + exp(-x)]$$

Where y is the output of the neuron and x is the total input to the neuron. The logistic sigmoid function limits the output y between 0 and 1.

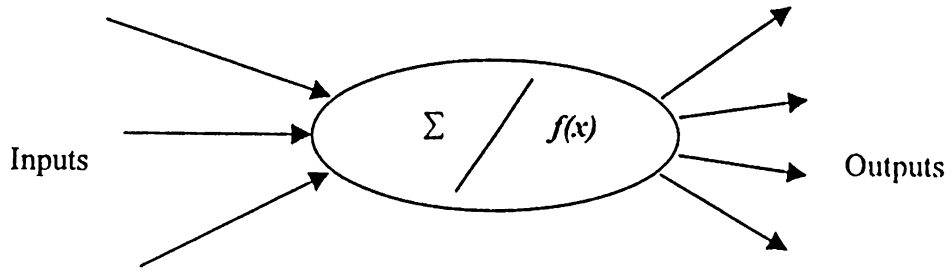


Figure 1. Formal neuron

The most popular and widely used form is the feed-forward NN trained with the back-propagation algorithm (Rumelhart *et al.*, 1986; Crick, 1989; Devillres, 1996). This specific network concurrent architecture is often called a back-propagation neural network (BPNN).

The BPNN is constructed from neurons (nodes) which are arranged in a series of layers. Each neuron is connected to others in the network by connection of different strengths or weights. The design of the network is the main feature influencing the flexibility of the model it generates. It is a number of neurons (nodes) in each layer, the number of layers, and the way the nodes are connected. The number of layers as well as well as the number of neurons in each layer depends on the application for which the NN model is designed and is, as a rule, determined by trial and error. The layers of neurons are usually fully connected.

Figure 2 shows a simple architecture of the NN system, which is based on input layer with 3 nodes, one-layer hidden units with 3 nodes, and one output layer with 2 nodes. The outputs from the input layer are multiplied by their weighting factors and added together as the input to a neuron in the next layer, the hidden layer. The output of the neurons from the hidden layer is calculated in the same fashion and propagated as interconnected inputs to the last layer of nodes called the output layer. The neurons of the output layer provide the actual output of the network, using the same procedure as is done by the neurons in the hidden layer. This process of computing output for a given input pattern is known as a feed-forward procedure.

The network is fully interconnected with their weighting factors. When any kind of signal is sent into the input layer, it is then propagated through the hidden layer towards the output layer. This is what is known as 'feed-forward' network, where the output of each layer is 'fed' into the inputs of the next layer in forward direction. What is called a layer of neurons is often described as a string, i.e. a linear arrangement of neurons (Medsker, 1994; De Wilde, 1997). Compared to standard statistical methods, three most important characteristics should be stressed:

1. There is no need to know the exact form of the analytical function on which the model should be built. This means that neither the functional type (polynomial, exponential, logarithmic, etc.) nor the number and positions of the parameters in the model-function need to be known.

2. All features known in standard model-generating techniques (choice of variables, methods for data and dimensionality reduction, experimental design, etc.) play an important role in BPNN procedure as well: the troublesome ones as well as the desirable ones.

3. The BPNN acts as a black box, allowing no physical interpretation of its internal parameters.

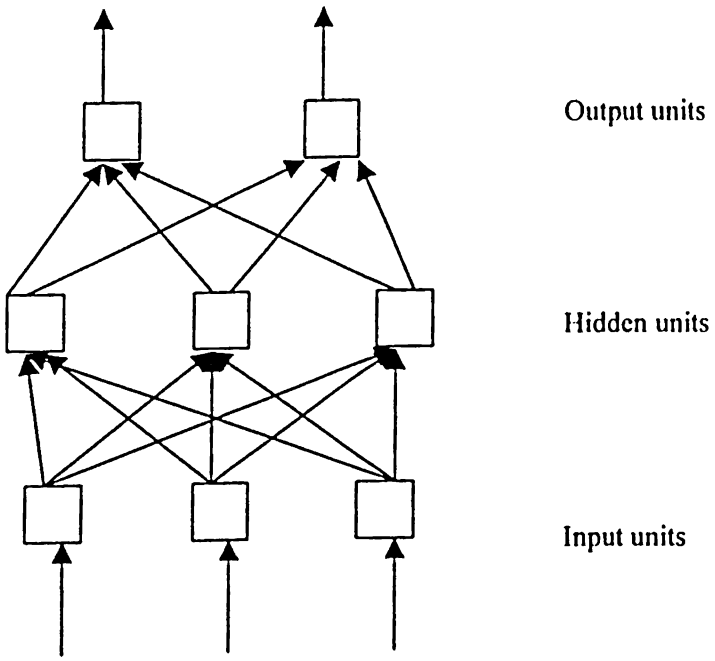


Figure 2. Classical BPNN architecture

The BPNN belongs to the class of supervised learning techniques, which means that the weights are corrected in order to produce 'correct' output values for as many inputs as possible. The method consists in comparing the response of the output units to the desired values via an iterative process in which an error term (SSE) is calculated and used to readjust the weights in the network. It requires the output (targets) to the inputs to be known in advance. The weights correction can be made after each individual new input (immediate correction), or after all inputs have been tested (deferred correction). Most applications use immediate correction.

The correction is made by generalized delta-rule. Learning by BP transports the data through the network in one direction and then scans through it, changing weights, in the opposite directions, using gradient descent method. This is an iterative least squares procedure in which the algorithm tries to adjust connection weights in a fashion which reduces the error most rapidly (Cartwright, 1993). An empirical learning constant rate (η) determines the speed of learning in an iterative procedure, i.e. changing the weights, while the inclusion of the momentum constant (μ) is necessary to avoid being trapped in a small local minimum. The momentum term has the power to prevent sudden changes in the direction towards the solution. Sum of squared errors from the output (SSE) must be also selected. For this design, the NN will be simulated and trained using Mathworks MATLAB software package, which uses total sum of the errors squared (SSE) as the convergence parameter. Once the network converges within the SSE criterion, the training process stops. Before the actual learning process starts, the following tasks must be satisfied:

- Initial choice of the NN model (the number of layers, neurons, and weights). It is only the starting point and the model is usually modified when more information is available from the results of learning and testing procedure.

- Randomization of initial weights by setting them to small random numbers.

- Selection of the learning rate η, momentum constant μ, SSE, and number of iterations (learning steps, epochs) for the training.

If the model does not converge to the desired SSE for the number of selected epochs, the systematic approach to improve the quality of the model is needed. It consists of the enlargement of the epochs (train the network longer), changing the network design, and expanding the training data set. But the most important aspect of designing a network for back-propagation learning is to ensure that it will not become overtrained. In that case, the rate of the error improvements slows down and the ability of the model to handle testing data extensively declines. As a result of the overtraining, all the outputs from the testing data are almost the same, called 'non-specific' outputs.

The most important feature of a BPNN is its ability to learn from examples and generalize, since the learned information is stored across the network weights. It is also able to make decisions and draw conclusions when presented with complex, noisy and partial information until some extent.

It is estimated that over 90 percent of the neural network applications in use today employ BPNN or some variant of it (Hammerstrom, 1995). Successful applications can be found in engineering (Hopfield, 1982; Kohonen, 1988; Freeman *et al.*, 1991), medicine (Miller *et al.*, 1992; Weinstein *et al.*, 1994), biotechnology and biochemistry (Montague *et al.*, 1994; Neal *et al.*, 1998), and chemistry (Zupan *et al.*, 1991; Tusar *et al.*, 1992; Schuster, 1992; Sumpter *et al.*, 1994; Kowalski *et al.*, 1995; Higgs *et al.*, 1997; Hervás *et al.*, 1998; Tetteh *et al.*, 1999).

In the same manner, the beginning of BPNN in quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) studies can be traced back to the early 1990s. This important part of neural network application for QSAR studies is discussed in the next chapter.

Collecting and preparing data

It is evident that data collection and analysis is a major part of the neural computing project. It runs often parallel with the other project activities because data is required in different quantities (data input vs. output) at different stages. Neural networks are data-driven models and the quality and efficiency of the model is greatly dependent on the quality of the data used to train it. Insufficient data (quality or the quantity) will stop the development of neural computing application from being successful. Data collection, preparation, analysis, and understanding are therefore essentials in the neural computing scheme.

Data requirements

In general, relevant and possibly pertinent data should be considered as inputs to the NN. It is not necessary to investigate the character of the relationship between the input and output (target) data, only that there is a strong possibility that there is a relationship. The NN will resolve an approximation to that function during the training process. The application is one of the modeling (prediction) or classification, learning will be supervised and target (output) data together with the input data must be collected. It is important to make a reasonable estimate of how much data is required to train the NN model properly. This is frequently in contrast with real applications since the amount of data (compounds in our case) is always limited. We need sufficient data and feature points for the form of the functional relationship and

to be specified accurately all through the whole input space for both, training and testing data sets. The NN cannot extrapolate reliably over a certain extent. If no training data are available in a region of input space from which some of the test data is drawn, then there is no valid generalization for those test patterns.

Data validation

Checking data validity can usually disclose any unacceptable values and errors. A simple validity check is a data range check. The following conditions should be checked and met before the input and output feature vectors are constructed.

1. All elements of the vector are within the expected range.
2. All vector elements are mutually consistent
3. The target (output) vectors are consistent with the input vectors. This is especially important for a supervised learning NN application.

Data partitioning

Data partitioning is a procedure for separating the data into training, testing, and validation sets. The overall set should contain enough data with a suitable distribution to guarantee that the NN model can learn the non-linear mapping and modeling between the input and output variables **x** and **y** over the whole range of feature vectors. In the first step, data should be allocated randomly for the whole data set. The training set should be sufficient to cover the multidimensional space of the whole data set with maximum diversity. The NN model accuracy will then be revealed in the procedure applied for the testing data set.

Design, training and testing of the NN prototype

NN model development is an iterative process based on systematic approach because the full complexity of the problem may not be instantly evident. It is not possible to create a full design specification initially since many design adjustments are implanted after some experimentation. Figure 3 shows the scheme of components, which are fundamental in the design procedure.

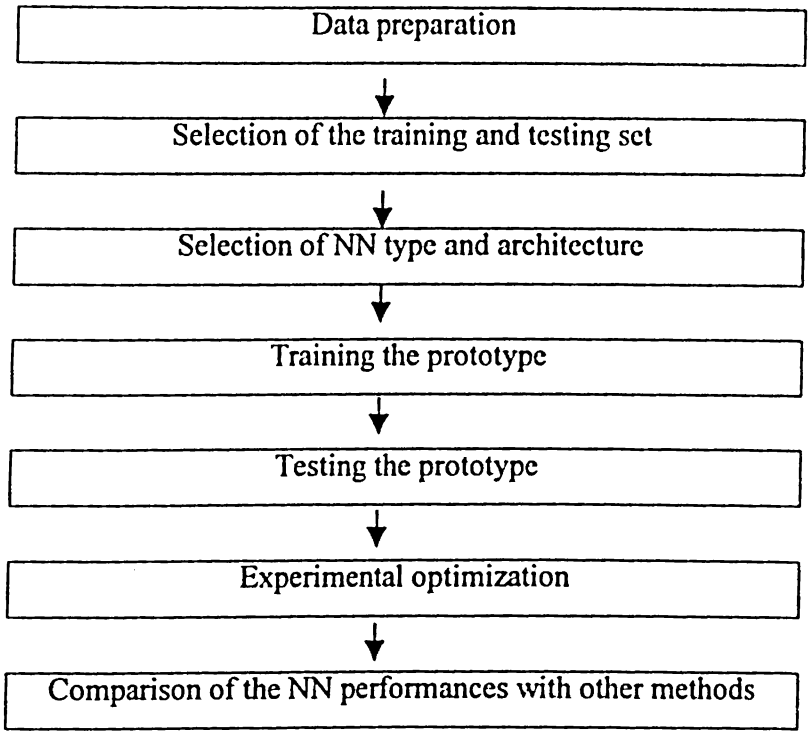


Figure 3. Components of the NN design

Data preparation

Neural computing systems very seldom operate directly on 'raw' data. Pre-processing the data set can often play a key role in the ability of the BPNN to carry out the desired mapping (Bos *et al.*, 1993). Data transformation can be used to fulfill the requirements imposed by the selected transfer function. The most commonly practiced is the min/max procedure, which transforms the data into <0,1> continuous interval according to the equation:

$$x^{T}_{i} = [(x_i - x_{min}) / (x_{max} - x_{min})]$$

Where x^{T}_{i} is scaled input, x_i the original input, x_{max} and x_{min} are respectively the maximum and the minimum values in each column of data matrix.

When the input data are multidimensional with high overlapping of different classes, the dimensionality must be reduced in order to limit the number of free parameters in the network and obtain a good generalization with the finite data set. To reduce the original pool of variables (descriptors) to an appropriate size, objective variable reductions should be performed using various criteria. The variables which contribute either no additional information or whose information is redundant with that of others already present in the system should be eliminated. What is required is a method to reduce the dimensionality of the data set while retaining its information content. Correlation matrix and multivariate factor analysis, for instance the principal component analysis (PCA) and cluster analysis, are suitable techniques to apply.

The idea of PCA is to find a small number of linear combinations of the original variables that maximize the variance accounted for in the original data. These new descriptors are called principal components, or latent variables, or properties (Marascuilo *et al.,* 1983). It leads to the eigensystem analysis of the covariance or correlation matrix. Spectral decomposition of covariance matrix is performed, then the eigenvectors and eigenvalues ordered. Eigenvectors corresponding to the largest eigenvalues are retained. Clustering analysis (CA) operates on calculated distances in multidimensional property space. Clustering can be applied to the clustering of compounds in a drug discovery project, clustering of substituent properties for the experimental design and clustering of entire database.

By reducing the number of inputs, the number of connections within the model is also reduced.

Selection of the training and testing set

The amount of data required for training a BPNN is dependent on the network architecture. The choice of a BPNN configuration must be made after numerous runs and frequently is a compromise between the learning and generalization abilities of the network. The performance of a BPNN during the training phase is evaluated by means of the error rate. The generalization capability is assessed from a testing phase. During this phase, unknown patterns are presented to the BPNN to check whether the prototype is able to predict their actual outputs. However, in practice, this condition is difficult to satisfy in QSAR studies where biological activities are measured only on

limited series of compounds. It is also important that, for the same data set, different solutions (i.e., models) exist. As a basic rule, at least 20% of the initial data set should be assigned to the testing set in order to train and test the network properly.

Selection of NN type and architecture

The number of hidden layers in BPNN depends on the complexity of the problem, but in most cases one hidden layer (Takahashi, 1993; Ito, 1994), with an optimal number of units, is sufficient. The problem is then reduced to the legitimate choice of number of hidden units. Numerous authors have proposed some empirical rules (Eberhart *et al.*, 1990; Leigh, 1995) but there is not any exact method to determine an appropriate network topology just from the inputs and outputs. It depends critically on the number of training cases, the complexity of the function or classification, and all variables and elements involved in this process. To establish the number of hidden nodes, which should be included into BPNN, a trial and error process is usually practiced. The approach consists in testing either bottom-up (adding more neurons if needed) or top-down combination (removing neurons, pruning the model).

Training the NN prototype

The back-propagation network is an iterative algorithm. Input vectors are presented to the system repeatedly (in random or non-random order), if the initial weights have to be updated after the introduction of each pattern. The output error (SSE) is calculated and the gradient descent algorithm is used to accommodate the weights and reduce the SSE in a back-propagation direction.

The major steps of the algorithm are as follows:

1. Initialize all the weights to random values.
2. Present a segment of input set and specify the desired target (output).
3. Output of all neurons is calculated using the present values of the weights in a feed-forward manner.
4. The total error for all nodes is computed (generalized delta rule).
5. The weights are adjusted by back-propagating the error information, where η is the learning rate, which controls the rate of adjustments of the weights.
6. The next segment is given to the system and the cycle is repeated until the total error term is equal to a given SSE criterion. Then the training process stops.

Testing the NN prototype

The test set is used to illustrate how well an already trained network can generalize by introducing previously unseen patterns to the network or system of networks. A common mistake is to select the best model on the basis of the test set performance. This is not a fair test of model accuracy, as the test set wrongly becomes part of the optimization procedure.

Experimental optimization

The optimal network performs best on the validation set (known outputs of the training set) with 'fixed' weights. Experimental optimization is the procedure when we allow the NN prototype longer training in order to improve the performance of the

existing model. If the operating conditions change over time, the testing data set is significantly different from the training data, or the feature vectors have been changed, the NN prototype has to be re-trained again. Once the performance of the NN prototype on the training and testing data is acceptable, the trained prototype could be transferred into the deliverable system. The obligation for maintaining confidence in the performance of the NN is to avoid the incorrect application of proposed NN model. The network system should not be allowed to extrapolate in all dimensions, i.e. it should not be used to determine features of input vectors where the majority of the data exceeds the degree to which different inputs are relevant, therefore the input vectors are outside the scope of the training input space.

Comparison of the NN performance with other methods

Comparison of the NN model performance with other methods is an important part of the prototype design and validates its efficiency. The performance of the optimal prototype is often substantial if compared with other multivariate methods such as principal component, discriminant, or regression analysis. Since the error term is summed over all output nodes for different models, the average SSE may not sufficiently measure the classification performances of a BPNN since it is linked to the value selected for a training threshold. It is then preferable to use another criterion. For the analogy of QSAR models obtained with a BPNN and other statistical tools, the percentages of good (or bad) classification can be employed, or the accuracy of different methods can be compared.

Statistical parametric methods are based on some assumptions. The input data are often assumed to be normally distributed (Bayesian classifier, De Wilde, 1997). But this is not true in many practical applications. Standard modeling techniques require the functional relationship to be known in advance. During the 'fitting' process, the parameters of this function are determined on the basis of the best agreement between the experimental input and calculated output data. The predictions are best when the experimental data cover the variable space evenly and with adequate density over the entire region without the extensive overlapping. The advantage of the NN model is that it does not require the knowledge of the functional association between input and output data sets. The non-linearity of a single unit transformation and a sufficiently large number of nodes and variable parameters (weights) guarantee enough 'freedom' to adapt the NN model to any existing relation between input and output data sets.

There is large overlay between the fields of neural networks and statistics. Statistics is concerned with data analysis. In neural network nomenclature, statistical interference means learning to generalize from noisy data. Most neural networks are not affected with data analysis, especially those intended to model biological systems, such as BPNN and therefore have little to do with statistics. But networks that can learn to generalize effectively from noisy data are similar to statistical methods. For example, a BPNN with no hidden layer is basically generalized linear model.

Common problems in training and testing the NN prototype

During the process of the NN design there are many problems which are commonly detected when training or testing the NN prototype. They can be divided into two fundamental categories based on the training or testing performances.

Poor training performance

Sometimes, the training error (SSE) will prevail continually high as the network fails to learn a mapping from the input set of feature vectors to the output data. The reasons for this abnormal response could be:

1. Incorrect choice of NN application, where there is an undetectable relationship between input and output data sets.
2. A functional relationship between both data sets exists but the selected features cannot properly describe this type of association.

Poor generalization performance

In this situation the error on the testing set is unacceptably high despite the successful training procedure. The reasons for this type of behavior could be:

1. Insufficient number of training patterns or very low diversity of training feature vectors. In this case there is no information in a region of input multidimensional space from which some of the test feature vectors are selected. Extrapolation over the reasonable extent will cause deficient generalization.

2. A network that is not sufficiently complex can fail to detect fully the signal in a complicated data set, leading to *under-fitting*. *Over-fitting* occurs when the complexity (number of weights) of the model is too high. In this case a network model may also fit the noise, not just a signal. Therefore, the first step to overcome this problem is to reduce the number of hidden units.

3. The model is trained longer (more epochs) but its ability to classify new data degenerates. This is known as the *over-training* effect. It can be explained as a consequence of parameter redundancy. The system has more parameters than are needed to solve a problem. The result is identical to *over-fitting* problem. The noise on the training patterns is learnt as well, leading to a decline in the generalization capacity on the testing set.

Proper designing of a BPNN is a systematic, frequently time-consuming process, which requires some experience and expertise to finally select the optimal model, since many elements must be selected, monitored, and adjusted.

Experimental design of neural network prototypes

Collecting data

Fifty taxol analog compounds were used to design a NN prototype. Data have been collected from the database of the National Cancer Institute, (NCI), Bethesda, MD 20892. Three different types of cancer: ovarian (OVA), breast (BRE), lung non-small cell (LNS) cancer, and an average across all 60 cell lines, represented by GI_{50} average values, are the principal targets (outputs) of the designing procedures. The central point is to reveal the quantitative structure-activity relationship (QSAR) between the molecular structure of a compound and its *in vitro* measured anticancer activity toward a specific type of cancer using a pattern recognition NN prototype.

Input data

The structure of the compound is the only source of input data. The features for each compound are calculated using CAChe software, version 3.1 (Computer-aided chemistry, Oxford Molecular Group, Inc., 2105 South Bascom Ave., Campbell, CA 95008; web page: http://www.oxmol.com/prods/cache). The software enables us to create a 'flexible' model of a molecular structure on a computer screen, and view the structure from many different angles and perspectives. The experiments in CAChe use mathematical models derived from computational chemistry to calculate molecular properties and geometry. The computation chemistry tools are derived from classical mechanics and quantum mechanics, and are applied to the chemical

sample by a number of computational applications that perform calculations based on specific parameters i.e. the entire array of calculations.

Using classical mechanics, CAChe can:

- optimize molecular geometry
- determine a series of energy conformations
- simulate the motion of atoms according to time, temperature and the calculated forces of atoms.

Using quantum mechanics, CAChe can:

- predict electron density and distribution in a molecule
- investigate molecular orbital energies
- optimize molecular geometry
- calculate molecular properties such as molecular weight, dipole moment, molar refractivity index, topological indices, atomic partial charge, energies and so on.

In this study, fifty molecular structures of taxol-analogs were drawn and saved in the library of compounds. The Cache software was then used to calculate all 27 chemical and physical properties (variables of the feature vector) for each compound. All the computed parameters are summarized in Table 1.

Parameter	Abbreviation
Molecular weight	MW
Atom count	AC
Bond count	DC
Conformation minimum energy	CME
Connectivity index of order 0	CI-0
Connectivity index of order 1	CI-1
Connectivity index of order 2	CI-2
Dipole moment	DM
Dipole vector X	DV-X
Dipole vector Y	DV-Y
Dipole vector Z	DV-Z
Steric energy	SE
Total energy	TE
Lipophilicity	LogP
Heat of formation	HOF
HOMO energy	HOMO
LUMO energy	LUMO
Molar refractivity	MR
Shape index of order 1	SI-1
Shape index kappa of order 1	SK-1
Shape index of order 2	SI-2
Shape index kappa of order 2	SK-2
Shape index of order 3	SI-3
Shape index kappa of order 3	SK-3
Valence connectivity index 0	VC-0
Valence connectivity index 1	VC-1
Valence connectivity index 2	VC-2

Table 1 Parameters of the feature vector

Based on computed values of every parameter for all the compounds, the input data set is composed of a (50 x 27) matrix. It means an array of numbers arranged into 50 rows (compounds) and 27 columns (variables).

Output data

The average activity (represented by negative $\log_{10}$ of GI_{50} molar concentration) for each type of cancer was calculated, while average values over all tested cell lines were obtained directly form the NCI web page (http://dtp.nci.nih.gov/).

Table 2 shows the cancer type with the cell lines used for the mean computation.

Cancer type	Cell line
Ovarian	OVCAR-4
	OVCAR-5
	OVCAR-8
	IGR-OV1
	SK-OV-3
Breast	MCF7/ADR-RES
	MDA-MB-231/ATCC
	MDA-MB-435
	MDA-N
Lung	NCI-H23
	NCI-H522
	A549/ATCC
	EKVX
	HOP-62
	NCI-H460

Table 2 Cell lines used in calculation of the activity average

Based on the cancer type, four different activities are executed (OVA, BRE, LNS, and average GI_{50}). Our goal is to design actually four neural network prototypes, where each one has to be optimized on the particular output that is an anticancer activity towards a specific cancer type. The outputs are four (50x1) matrices, each representing a distinct type of cancer.

Therefore, it is important to make clear that for all NN prototypes, the input data are fixed (50 compounds and 27 properties), while the output (activity) differs on the type of cancer cell line.

Data preparation

Both data sets, input and output data must be rescaled from their 'raw' numerical style according to the requirements of the activation function. Since the logistic sigmoid function has been used, the transformation into the <0,1> continuous interval is relevant. The min/max formula is applied for each parameter (column) of input data. Since the variance of output data is very modest, all the outputs are at that point converted into one <0,1> analog interval. This procedure provides the opportunity to compare the NN prototypes for different types of cancer. The last step of data preparation is to allocate the data randomly for the whole data set. The random numbers are generated and the data set is sorted, based on the appropriate random digit. Consequently, all compounds remain in this 'fixed' random arrangement during the entire systematic process of designing BPNN prototype.

Selection of the training and testing set

The first 40 compounds in random order (80%) are assigned to the training (learning) set, while the rest of the data; i.e. 10 compounds (20%) are reserved for the testing set. This is the compromise between the learning and generalization capabilities of the NN model. The training set should sufficiently cover the multidimensional space

of the entire data set with maximum diversity. On the other hand the biological activities are always measured on the limited set of compounds. The validation set, comparing the accuracy of the model, called 'compa', contains the output (activity) of all 50 compounds.

Selection of the NN type and architecture

The NN is simulated, trained, and tested on the MATLAB software, (version 5.3.0. 1999). Feed-forward back propagation neural network (BPNN) utilizing gradient descent method with one hidden layer and flexible number of units together with one output node (activity) is used. The number of input nodes is also a part of the model systematic design and depends on the multidimensionality reduction analyses, i.e. analysis of the correlation matrix, principal component analysis (PCA), and pattern analysis.

Input, output, test and compa data could be read directly from a Microsoft Excel (filename.xls) through the 'Read procedure' into the MATLAB software. For the initiative network topology, the training parameters (TP) for the program have to be selected together with the initial randomization of the weights.

TP= [displayed frequency; maximum iteration (epochs); SSE; learning rate (η); learning rate increment; learning rate decrement; momentum constant (μ); error ratio]

Here is an example of the MATLAB program, using the feed-forward back propagation neural network (BPNN):

```
>> diary filename;
>> diary on                          % Creating the working  diary textfile

>> clear;                            % Clear all variables
>> load data;                        % Input, output, test, and compa data are read

>> In = input';                      % Data matrix (27 x 40)
>> Out = output';                    % Data matrix (1 x 40)
>> Test = test';                     % Data matrix (27 x 50)
>> Compa = compa';                   % Data matrix (1 x 50)

>> S1 = 27;                          % Number of input variables
>> S2 = 27;                          % Number of flexible nodes in the hidden layer
>> S3 = 1;                           % Number of output units

>> Seed = 11;                        % Randomization of initial weights
>> rand('seed', Seed);
>> [w1, b1] = rands(S1, S1);
>> [w2, b2] = rands(S2, S1);
>> [w3, b3] = rands(S3, S1);

>> TP = [200  50000  0.01  2  1.05  0.7  0.7  1.04];  % Training parameters
>> f = 'logsig';                     % Activation sigmoid function

>> [w1,b1,w2,b2,w3,b3,TE,TR] = trainbpx(w1,b1,f,w2,b2,f,w3,b3,f,In,Out,TP);
```

The NN is now trained and number of epochs, learning rate, and SSE displayed. When NN reaches the selected SSE (0.01), the training process stops. Then the testing procedure starts.

```
>> [A1, A2, A3] = simuff(Test, w1, b1, f, w2, b2, f, w3, b3, f );
>> A3                                % The generalized (predicted) output
>> A3 – compa                        % The errors are presented in the same order

>> diary off                         % Diary text-file is saved
>> save model
```

Setting for the model (last propagation) with all the calculated variables and adjusted weights is saved in the MATLAB file.

Analysis of the prediction accuracy

By calculating the differences (Δ) between the prototype predictions and the actual values of the activities, represented in the model by (A3 – compa), the criterion for the model accuracy must be specified. The tested hypothesis is:

Is the antitumor activity (output) of this compound against the particular cell line higher or the same as taxol activity?

Prediction of the model is considered to be erroneous when the predicted value differs from the actual value more than ± 0.1.

It means: $\text{Error} = \Delta = | A3 - \text{compa} | > 0.1$

The chosen error criterion is actually non-linear since it is assigned to already scaled data. The output data (activities) were scaled according to min/max procedure with the equation:

$$x^T_i = [(x_i - x_{min}) / (x_{max} - x_{min})]$$

Where the $x_{min} = 4.00$ and $x_{max} = 10.00$ represent the negative logarithms of the GI_{50} molar concentration. For complete understanding of the model error, the data must be unscaled to their raw value according to the simple back-transformation equation:

$$x_i = 6x^T_i + 4$$

This process is illustrated for taxol compound (identification number NSC 125973) with activity towards ovarian cancer measured on the GI_{50} scale as 0.616.

The error $\Delta = \pm 0.1$ is applied in the following way:

1. No error equals the activity of 2.01×10^{-8} M

   $x_i = (0.616)6 + 4 = 7.696$, then the concentration $10^{-7.696} = 2.01 \times 10^{-8}$ M

2. Positive error (+0.1) equals the activity of 0.51×10^{-8} M

   $x_i = (0.616 + 0.1)6 + 4 = 8.296$, then the concentration $10^{-8.296} = 5.1 \times 10^{-9}$ M

3. Negative error (-0.1) equals the activity of 8.02×10^{-8} M

   $x_i = (0.616 - 0.1)6 + 4 = 7.096$, then the concentration $10^{-7.096} = 8.02 \times 10^{-8}$ M

While the error is included in the calculation, the activity interval for this particular compound is in the range of $(0.51, 8.02) \times 10^{-8}$ with the value of 2.01×10^{-8} as the true activity. This example shows the non-linearity of the selected error, which is used as a criterion to evaluate the BPNN model accuracy.

Dimensionality reduction

Initial BPNN with all variables does not converge and reveals very poor training and generalization performances. Dimensionality reduction is the next logical step to enhance the model performance. To reduce the original number of variables (27) in the input set, analysis of the correlation matrix, pattern, and principal component analyses were performed.

Correlation matrix is a square, diagonally symmetric arrangement of correlation coefficients that measure the strength of the linear relationship between two variables. Spearman rank correlation was calculated by StatMost computer software (Statistical analysis and graphic, version 2.50, DataMost Corp., Salt Lake City, UT 84164), because it is a nonparametric statistic that does not require the assumption of normality and constant variance of the residuals. Since the matrix dimension is (27x27), only a part of that matrix with highest Spearman correlation coefficients (SCC > 0.95) is presented in Table 3.

Variable	Variable	SCC	Variable	Variable	SCC
BC	AC	0.994	SI-1	CI-0	0.991
CI-0	MW	0.970	SI-1	CI-2	0.952
CI-1	MW	0.957	SI-1	TE	0.966
CI-1	CI-0	0.956	SI-1	HOF	0.966
CI-2	MW	0.957	SK-1	TE	0.971
CI-2	CI-0	0.974	SK-1	HOF	0.971
SE	CME	0.973	SK-1	SI-1	0.956
TE	MW	0.964	SI-2	CI-1	0.958
TE	CI-0	0.965	SI-2	MR	0.956
HOF	MW	0.964	SI-3	SI-1	0.951
HOF	CI-0	0.965	VC-0	TE	0.952
HOF	TE	1.000	VC-0	HOF	0.952
MR	MW	0.950	VC-0	SK-1	0.971
MR	CI-1	0.967	VC-1	SK-2	0.951
SI-1	MW	0.961	VC-1	VC-0	0.958

Table 3 Analysis of correlation matrix

The suggestions based on analysis of the correlation matrix are as follows:

1. While there is a total correlation between heat of formation (HOF) and total energy (TE), the variable with the lower variance can be deleted from the data set.

2. Most of the topological parameters, i.e. connectivity (CI), shape (SI, SK), valence connectivity indexes (VC), molecular weight (MW), atom count (AC), and bond (BC) count are highly correlated.

3. There is not a significant correlation for the CME, DM, SE, LogP, HOMO and LUMO energies, and MR.

4. Reducing the dimensionality of the data set, based entirely on the correlation analysis, could be a misleading procedure at times since even highly correlated variables could still be interesting from the vantage point of pattern and other analyses.

Principal component analysis (PCA) was carried out using the same (StatMost) software. The eigenvalues are an index of precision, therefore the larger eigenvalue the better the precision in the direction of the corresponding eigenvectors (principal components). The spectral decomposition (factorization) was performed on the covariance matrix of the input data. The data are summarized as a linear combination of an orthonormal set of vectors. Table 4 shows the results of the PCA.

Variable	Eigenvalue	Proportion	Cumulative
MW	17.25	0.784	0.784
DM	2.31	0.105	0.889
BC	1.1	0.049	0.938
CME	0.7	0.032	0.969
CI-0	0.4	0.016	0.986
CI-1	0.17	0.01	0.993
CI-2	0.07	0.003	0.997
SE	0.04	0.002	0.998
LogP	0.02	0.001	0.999
HOF	0.01	0.0004	1.000

Table 4 Principal component analysis

The results of the PCA show that those 10 variables have the cumulative variance of 100% based on the eigenvalues.

Pattern analysis reveals the parameters, which are fundamental in pattern recognition procedure, no matter how they are correlated and what their variance is. Without running NN, data of each variable are resorted upon their recognition capacity. Starting with only three variables (LogP, DM, and SE) sorted by descending order and comparing their patterns towards the binary value of activity, the same patterns leading into different outputs could be easily analyzed. The next variable is then selected to solve the consequent pattern uncertainty, which means to reduce a number of wrong classifications. The results of the procedure of adding feasible variables are shown in Table 5.

Number of variables	Number of wrong classification
3	16
4	14
5	11
6	10
7	9
8	9
9	7
10	10
11	11
12	11

Table 5 Pattern analysis

Pattern analysis displays the unusable variables for the NN model as well. For example, the variable HOMO energy has no pattern variability since the variance is extremely small and all the raw data are in the interval range (-11.927, -11.360). This is true for the LUMO energy as well, since all the data remain within a very small cluster.

Based on the results of all the previously performed analyses the initial dimension of input data set with 27 variables could be efficiently reduced to 9 final parameters without any serious loss of information. The reduction is very substantial, since this means the number of the connections within a model is also reduced. Table 6 displays the final parameters used in BPNN prototype design (variable S1) with the scaled input data subset for the first 10 compounds.

#	LogP	DM	DV-X	DV-Y	DV-Z	SE	CME	HOF	MR
1	.3724	.0496	.7873	.2675	.2044	.0192	.0335	.0647	.5243
2	.3977	.0766	.5667	.2720	.1009	.1993	.8520	.0724	.4849
3	.7487	.0544	.6481	.3040	.1895	.0169	.0399	.0384	.6023
4	.8486	.0545	.6493	.2423	.1155	.0207	.0498	.0459	.6360
5	.4501	.1134	.7653	.2762	.0000	.8977	.9444	.0496	.5590
6	.6789	.0297	.6294	.2754	.1764	.0153	.0292	.0568	.5865
7	.9450	.1112	.4000	.2882	.1348	.3649	.6294	.9991	.6446
8	.6451	.0480	.7644	.2391	.2077	.0194	.8968	.0267	.6482
9	.6995	.2641	1.000	.2609	.5382	.0189	.0454	.9936	.6546
10	.9353	.1506	.8442	.2147	.3956	.0125	.0453	.9935	.6856

Table 6 Subset of the input data for 10 compounds

Selecting the number of feature vectors

The BPNN was initially run on only 2 feature vectors (compounds) with maximum difference in the output values. Then another pair of vectors was added in the same manner. The procedure continues by summing up another couple of vectors until all 40 out of 50-feature vectors are used. The results of the optimization are summarized in Table 7. In each NN run, all previously established variables (9) were used as the input data. The results indicate that all training compounds (40) are useful in the BPNN prototype design, since the error rate for the full set is minimal. This selection process is done only on the ovarian (OVA) input data and then successfully used on other cancer types (including the average GI_{50}).

Error/10 is the error (Δ>0.1) from the testing set, while Error/50 stands for the total accuracy of the model represented by all (50) compounds in the validation set.

Both errors show a decreasing rate as the number of feature vectors increases.

Feature vectors	Number of epoch	SSE	Error/10	Error/50
2	55	0.007	8	39
4	48	0.008	9	37
6	58	0.010	9	41
8	252	0.010	8	34
10	540	0.010	10	35
12	9187	0.010	10	36
14	9809	0.010	8	31
16	50000	0.308	6	24
18	50000	0.032	6	23
20	50000	0.023	7	21
22	50000	0.022	5	18
24	50000	0.023	8	23
25	50000	0.023	8	23
26	50000	0.023	6	18
28	50000	0.014	5	14
30	50000	0.059	5	13
32	50000	0.065	5	12
34	50000	0.028	9	14
36	50000	0.021	5	10
38	50000	0.029	6	9
40	**50000**	**0.070**	**3**	**7**

Table 7 Optimization the number of feature vectors

Selecting other parameters

After the assortment of the NN learning rate, momentum variables, and feature vectors for the model, selection of the number of hidden units, epochs, and initial weights is the next important step in the design procedure. There are no exact methods to determine an optimal network topology just from the inputs and outputs. The trial and error process is used with the results presented in subsequent tables.

The BPNN is trained until it reaches the selected SSE (0.01) or the limit of training epochs. There is one BPNN prototype for each cancer type, including the GI_{50} output. Table 8 represents the design for OVA and LNS, while Table 9 illustrates the design for BRE cancer and the GI_{50} index.

Cancer type	Hidden nodes (S2)	Number of epoch	Weights (Seed)	SSE	Error/10
Ovarian	0	50000	11	1.793	7
	0	25000	11	1.382	7
	1	10000	11	0.146	4
	1	**50000**	**11**	**0.070**	**2**
	1	70000	11	0.013	4
	1	100000	11	0.022	4
	1	50000	7	0.079	4
	1	50000	13	0.029	8
	2	34259	11	0.010	5
	3	35000	11	0.017	5
	3	50000	11	0.162	6
	9	50000	11	0.054	7
	18	50000	11	0.113	7
Lung	0	50000	11	1.931	7
	1	50000	3	0.052	7
	1	50000	10	0.051	6
	1	**50000**	**11**	**0.062**	**3**
	1	50000	12	0.024	6
	1	50000	13	0.059	7
	1	50000	17	0.035	10
	1	50000	35	0.056	6
	2	47258	7	0.010	6
	2	50000	11	0.013	6
	2	53369	15	0.010	5
	3	50000	11	0.057	6
	5	50000	11	0.050	7
	9	50000	11	0.051	8
	10	50000	11	0.056	9
	18	50000	11	0.052	7

Table 8 BPNN design for OVA and LNS cancer

Cancer type	Hidden nodes (S2)	Number of epoch	Weights (Seed)	SSE	Error/10
Breast	0	50000	11	0.962	5
	1	50000	3	0.054	6
	1	50000	5	0.044	7
	1	50000	7	0.107	7
	1	50000	10	0.113	7
	1	50000	11	0.051	6
	1	50000	12	0.041	5
	1	**50000**	**13**	**0.024**	**3**
	1	69288	13	0.010	5
	1	52442	15	0.010	5
	1	50000	17	0.051	7
	2	50000	7	0.071	7
	2	50000	11	0.057	6
	2	50000	13	0.062	7
	2	50000	15	0.043	6
	3	50000	11	0.050	6
	9	50000	11	0.061	8
	18	50000	11	0.057	7
GI_{50}	0	50000	11	1.438	6
	1	50000	3	0.025	6
	1	50000	5	0.032	7
	1	50000	7	0.044	4
	1	50000	10	0.035	5
	1	50000	11	0.048	4
	1	50000	12	0.023	5
	1	**50000**	**13**	**0.060**	**3**
	1	50000	15	0.027	6
	1	50000	17	0.026	7
	2	25291	7	0.010	7
	2	50000	11	0.027	7
	2	50000	13	0.051	8
	2	23773	15	0.010	5
	3	50000	11	0.332	6
	9	50000	11	0.046	7
	18	50000	11	0.037	7

Table 9 BPNN design for BRE cancer and GI_{50}

Bolded rows in the tables express the best prototype design based on the generalization capability of the model and number of errors (Δ>0.1) for testing set. The results, based on the same number if learning epochs (50000 iterations), together with both types of errors are summarized in the next Table 10.

Cancer type	Hidden nodes (S2)	Number of epoch	Weights (Seed)	SSE	Error/10	Error/50
Ovarian	1	50000	11	0.070	2	6
Lung	1	50000	11	0.062	3	5
Breast	1	50000	13	0.024	3	3
GI_{50}	1	50000	13	0.060	3	5

Table 10 Best BPNN prototypes

Experimental optimization

In this stage of BPNN design, the training process (number of epochs) is allowed to become the only variable under investigation. Other parameters (S1, S2, and randomization of the initial weights) are kept in their rigid configuration, i.e. S1 = 9, S2 = 1, and Seed = 11 or 13. This procedure is often called a vertical analysis (Devillers, 1996). The results are presented in Table 11 (for OVA and LNS) and Table 12 (for BRE and the index GI_{50}).

Bolded rows in the tables indicate the optimized prototype based on the generalization capability of the model, minimal number of errors (Δ>0.1) for both sets (test and validation), and minimum of sum of squared errors, i.e. $\Sigma\Delta^2$ = min.

Cancer type	Number of epoch	SSE	Error/10	Error/50
Ovarian	10000	0.137	5	11
	20000	0.095	4	8
	30000	0.080	3	7
	40000	0.074	2	6
	45000	0.072	2	6
	50000	0.070	2	6
	55000	0.068	2	6
	60000	0.064	2	5
	65000	**0.057**	**2**	**4**
	70000	0.049	4	5
	100000	0.022	4	9
Lung	10000	0.170	5	10
	20000	0.116	3	6
	30000	0.098	3	6
	35000	0.092	3	6
	40000	0.086	3	6
	45000	0.077	3	6
	50000	0.062	3	5
	55000	0.047	3	4
	60000	0.041	3	4
	65000	0.038	3	4
	70000	0.035	3	4
	75000	0.034	3	4
	80000	**0.032**	**3**	**4**
	85000	0.031	4	6
	90000	0.029	5	7

Table 11 Optimized BPNN for OVA and LNS cancer

Cancer type	Number of epoch	SSE	Error/10	Error/50
Breast	10000	0.107	6	9
	20000	0.076	7	9
	30000	0.050	6	8
	40000	0.034	5	6
	45000	0.028	3	4
	50000	0.024	3	3
	55000	**0.019**	**3**	**3**
	60000	0.015	3	4
	65000	0.012	4	4
	69288	0.010	5	5
GI_{50}	5000	0.251	7	14
	10000	0.152	4	8
	15000	0.113	4	7
	20000	0.094	4	7
	25000	0.082	4	6
	30000	0.074	4	6
	35000	0.067	4	6
	40000	0.062	4	6
	45000	0.060	4	6
	50000	0.060	3	5
	55000	0.057	3	5
	60000	0.058	3	5
	65000	**0.055**	**3**	**5**
	70000	0.054	3	5
	80000	0.054	3	5
	100000	0.053	4	7

Table 12 Optimized BPNN for BRE cancer and GI_{50}

Table 13 exhibits the results of BPNN generalization capacity, based on the optimal learning epochs (iterations), together with both types of errors (Δ) for the efficacy validation.

Cancer type	Hidden nodes (S2)	Number of epoch	Weights (Seed)	SSE	Error/10	Error/50
Ovarian	1	65000	11	0.057	2	4
Lung	1	80000	11	0.032	3	4
Breast	1	55000	13	0.019	3	3
GI_{50}	1	65000	13	0.055	3	5

Table 13 Optimal BPNN prototypes

Comparing the results displayed in Table 10 and Table 13, there is a valuable enhancement in the NN performance (accuracy of the model) for the type of ovarian and lung cancer. The errors in the validation set decrease by 2 and 1, respectively. The model accuracy is improved from the original 80% to 92% in case of ovarian cancer and from 90% to 92% for lung cancer. SSE value is improved (reduced) from the initial (0.024, 0.070) to optimal (0.019, 0.057) but the initial SSE criterion (0.01) has not been reached. This implies that further training will most likely approach the selected SSE criterion. However, the BPNN prototype will become probably overtrained with non-specific tested outputs.

Optimal BPNN models are saved in MATLAB files (filename.mat), therefore, they are ready to be used for the predictive competence of the prototype without any need to design or train the model a second time.

Comparison of the NN performance with other methods

The comparison of the NN prototype performance with other methods validates its efficacy. As the comparative methods, the multivariate regression analysis and Bayes' rule analysis are used.

The multiple regression analysis results are actually created by the NN model with no hidden nodes. These data are already generated during the model design procedure.

The Bayes' rule (De Wilde, 1997) is a discriminant analysis based on the assumption of data multivariate normal distribution. Output data are binary allocated to the different classes. In our study, class 1 represents the inputs that lead to the output activity equal or higher than taxol anticancer activity. Class 0 then designates the subset of compounds with their activity lower than taxol. The analysis is based on the calculation of the average vectors of each class, using the discriminant function and then comparing the Mahalanobis (or Euclidean, in case of the singular matrix) distances of the sample to each average. When the distance is smaller, the sample is assigned to that particular class.

This approach, with only binary outputs, is frequently used, since the BPNN prototype results can be easily transformed into the binary outputs as well. Table 14 shows the results of PBNN compared to other methods with both, testing and validation errors (Δ).

Cancer type	Optimal BPNN (binary) E/10, E/50	Multivariate Regression E/10, E/50	Bayes' rule E/10, E/50
Ovarian	2, 3	7, 33	7, 21
Lung	3, 3	7, 28	3, 16
Breast	1, 2	5, 19	2, 12
GI_{50}	3, 4	6, 24	4, 20

Table 14 Comparison of the BPNN with other methods

Table 14 shows higher efficacy of the BPNN model (lower number of errors) over other parametric statistical methods. The BPNN results were converted to the binary outputs; hence, the comparison with the Bayes' analysis is a valid procedure.

For breast cancer, the Bayes' analysis seems to have very good generalization capacity for the testing set (2 errors only), but when all the compounds in the validation set are tested, the method gives about 76% accuracy, while the BPNN is 96% accurate.

Statistical techniques are not adaptive, but typically process all training data simultaneously. Neural network classifiers are non-parametric systems that make weaker assumptions concerning the shapes of underlying distributions. Therefore, they are more robust when distributions are generated by nonlinear processes, and are strongly non-Gaussian. The application of the NN does not require detailed knowledge of the transformation complex function of the measured system.

However, we should keep in mind that NN is just a powerful tool that can help us solve some specific problems, and for many other applications, the standard parametric methods of data fitting and approximation techniques could be even better than NN in the sense of the overall performance in model accuracy, with a lower number of adjustable parameters.

Results

The results, in terms of model accuracy, are based on generalization capability of the optimal prototype. Prediction of the model is considered an error, providing the predicted value differs from the actual value more than ± 0.1. The error criterion was selected as a variable of the BPNN architecture design. The non-linear characteristic of this parameter was documented in previous text. Table 15 illustrates the accuracy for every **analog** BPNN optimal prototype together with multiregression analysis.

Cancer type	BPNN Tested accuracy %	BPNN Total accuracy %	Regression Tested accuracy %	Regression Total accuracy %
Ovarian	80.0	92.0	30.0	34.0
Lung	70.0	92.0	30.0	44.0
Breast	70.0	94.0	50.0	62.0
GI_{50}	70.0	90.0	40.0	52.0

Table 15 Accuracy of the analog BPNN and multiregression

Table 16 shows the accuracy for each **binary** BPNN optimal prototype and the Bayes' discriminant analysis.

Cancer type	BPNN Tested accuracy %	BPNN Total accuracy %	Bayes' rule Tested accuracy %	Bayes' rule Total accuracy %
Ovarian	80.0	94.0	30.0	58.0
Lung	70.0	94.0	70.0	68.0
Breast	90.0	96.0	80.0	76.0
GI_{50}	70.0	92.0	60.0	60.0

Table 16 Accuracy of the binary BPNN and Bayes' rule

When different techniques are used, correlation between predicted and actual outputs is often accepted as an additional accuracy criterion. Unconventional modeling techniques and designs could be easily compared, based on their results, instead of pre-selected error estimation. The correlation coefficient describes the strength of the linear relationship between predicted and measured activities. Table 17 reveals calculated correlations of validation data set for each optimal PBNN prototype.

Cancer type	Correlation (predicted vs. measured activity)
Ovarian	0.831
Lung	0.945
Breast	0.913
GI_{50}	0.886

Table 17 Correlation between predicted and real activities

A comparative molecular field analysis (CoMFA) was performed on the subset of taxol analogues (Czaplinski *et al.*, 1994) with correlation 0.853 between predicted and real activities. It is evident, by comparison of the correlation coefficients of both techniques, that BPNN can achieve even more accurate predictions in case of lung and breast cancer types and index GI_{50}.

Conclusion

The back-propagation neural network (BPNN) is a powerful technique in QSAR and QSPR studies. In fact, a BPNN can find generally non-linear relationships between the structure of the molecules and their activity or property. Fundamentally, the BPNN can learn from examples and can gain its own knowledge (the model is trained). It's able to solve many new problems by performing generalization. BPNN is also error tolerant and can handle a certain level of noisy or incomplete (corrupted) data. However, it presents some drawbacks. The designing and training phase is time-consuming. Applying the BPNN requires some adjustments of many parameters (e.g. data pre-processing, architecture, SSE, learning rate, momentum and so on), that strongly control its behavior. A large amount of effort is required to overcome these problems. They basically deal with some modifications of the original BPNN algorithm and the creation of a hybrid system.

The BPNN is definitely not a cure-all recipe in QSAR studies, but this very potent tool cannot be ignored in the domain of structure-activity relationship and should be used parallel with other classical methods.

References

Bos, M.; Bos, A.; Linden, W.E. Data processing by neural networks in quantitative chemical analysis. *Analyst*, **1993,** *118*, 323-328.

Cartwright, H.M. Applications of artificial intelligence in chemistry, Oxford University Press Inc., New York, **1993,** 13-39.

Crick, F. The recent excitement about neural networks. *Nature,* **1989,** *337,* 129-132.

Czaplinski, K.H.; Grunwald, G.L. A comparative molecular field analysis (CoMFA) derived model of the binding of Taxol analogues to microtubules. *Bioorganic & Medicinal Chemistry Letters,* **1994,** *4 (18),* 2211-22216.

Devillers, J. Genetic algorithms in computer-aided molecular design. In, *Genetic algorithms in molecular modeling* (J. Devillers, Ed.). Academic Press, London, **1996,** 1-34.

De Wilde, P. Neural network models: an analysis. Springer-Verlag London Limited, second edition, **1997,** 33-48.

Eberhart, R.C.; Dobbins, R.W. Implementations, In, *Neural Network PC Tools. A practical guide.* (R.C. Eberhart and R.W. Dobbins, Eds), Academic Press, San Diego, **1990**, 35-58.

Freeman, J.A.; Skapura D.M. Neural networks: Algorithms, applications, and programming techniques. Addison-Wesley Publishing Co., Reading, MA, **1991.**

Hammerstrom, D. A digital VLSI architecture for real-world applications. In, *An introduction to neural and electronic networks*, Second Edition, Academic Press, San Diego, **1995**, 335-358.

Hervás, C.; Venture, S. Computational neural networks for resolving non-linear mulicomponent systems. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 1119-1124.

Higgs, R.E.; Bemis, K.G.; Watson, I.A.; Wikel, J.H. Experiment design for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci.*, **1997**, *37*, 861-870.

Hopfield, J.J. Neural networks and physical systems with emergent collective computation abilities. *Proc. Natl. Acad. Sci. USA*, **1982**, *79*, 2554-2558.

Ito, Y. Approximation capability of layered neural networks with sigmoid units on two layers. *Neural Comput.*, **1994,** *6,* 1233-1243.

Kohonen, T. An introduction to neural computing. *Neural Networks*, **1988,** *1(1),* 3-16.

Kowalski, B.R.; Wang, Z.; Hwang, J.N. ChemNets: Theory and application, *Anal. Chem.*, **1995,** *67*, 1497-1504.

Leigh, D. Neural networks for credit scoring. In, *Intelligent systems for finance and business.* (S. Goonatilake and P.Treleaven, Eds.) John Wiley & Sons, Chichester, **1995,** 61-69.

Marascuilo L.J.; Levin, J. Multivariate statistics in the social sciences, Brooks/Cole Publishing Company, **1983,** Chapt.6.

Medsker, L. Neural network connections to expert systems. In, *World Congress on neural networks.* San Diego, Lawrence Erlbaum Associates and INNS Press, **1994,** 411-417.

Miller, A.S.; Blott, B.H.; Hames, T.K. Review of neural network applications in medical imaging and signal processing. *Med. Biool. Eng. Comput.*, **1992,** *30*, 449-464.

Montague, G.; Morris, J. Neural-network contributions in biotechnology. *TIBTECH*, **1994,** *12,* 312-324.

Neal, M.J.; Gooddarce, R.; Kell, D.B. Toward generating neural network structures for functional approximation. *Neural Networks*, **1998,** *11,* 89-99.

Pao, Y.H.; Adaptive pattern recognition and Neural Networks. Addison-Wesley Publishing company, **1989,** 309.

Ripley, B.D.; Statistical aspects of neural networks. *J. Chem. Inf. Comput. Sci.*, **1993,** *33,* 202-210.

Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature,* **1986,** *233,* 533-36.

Schuster, H.G.; Applications of neural networks. VCH Publishers Inc., New York, **1992,** 153-239.

Sumpter, B.G.; Getino, C.; Noid, D.W. Theory and applications of neural computing in chemical science. *Ann. Rev. Phys. Chem.*, **1994,** *45,* 439-481.

Takahashi, Y. Generalization and approximation capabilities of multilayer networks. *Neural Comput.*, **1993**, *5*, 132-139.

Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. Quantitative structure-property relationships for estimation of boiling point and flash point using a radial basis function neural network. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 491-507.

Tusar, M.; Zupan, J.; Gasteiger, J. Neural networks and modeling in chemistry. *J. Chim. Phys.*, **1992**, *89*, 1517-1529.

Wasserman, P.D. Neural computing: theory and practice. Van Nostrand Reinhold, New York, **1989**, 255.

Weinstein, J.N.; Myers, T.; Casciari, J.J.; Buolamwini, J.; Raghavan. K. Neural networks in the biomedical sciences: A survey of 386 publications since the beginning of 1991. In, *World Congress on neural networks*, San Diego, Lawrence Erlbaum Associates and INNS Press, **1994**, 121-128.

Zupan, J.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta*, **1991**, *248*, 1-30.

Zupan, J.; Gasteiger, J. Neural networks for chemists – an introduction. VCH Publishers, New York, **1993**, 9-36.

Chapter 4

QSAR OF TAXOL ANALOGUES USING NEURAL NETWORK PROTOTYPES

Abstract

The neural network (NN) prototype is used as a powerful tool for the design of quantitative structure-activity relationships (QSAR), screening of structurally similar taxol analogues for their anticancer activities, and a significant prediction of potential pharmaceutical agents. Predictions are made within a selected error ($\Delta = \pm 0.1$). A test set of taxol analogues is subjected to a neural network model trained on calculated fundamental physicochemical descriptors and on anticancer activity data. The prediction of the activity is based on the generalization performance of the NN prototype. Taxol analogues are then sorted and presented in tabulated forms to determine the compounds with the highest activities against a particular kind of cancer. Compounds with the highest predicted antitumor activities, 10y110905, 10y110938, and 10y110963, are taxol analogues with a substituent propionic, 1-methyl-2-pyrrolecarboxylic, and crotonic acid in position C-10 (10-O-analogues). These compounds are the best candidates for anticancer screening and might be expected to verify the predicting power of the neural network prototype.

Introduction

Mankind has been suffering from cancer since antiquity. Cancer is an illness based on abnormal cell growth and development. The result is the formation of a new tissue, cancer tissue. Physicians and scientists have begun to learn and understand these errors in cell growth. It is known that there is more than one error involved in cancer development. Since we are living longer, cells are actually exposed to a higher risk of developing such errors. As a result, there are more cancer patients now as compared to 50 or 100 years ago. There is also a significant difference in cancer rates in countries where the life span is shorter than in the United States. Americans above the age of 65 years have an almost tenfold risk of developing cancer than their younger counterparts. The most common forms are breast, colon, lung, and ovarian cancers.

Breast cancer is the most common cancer in women, being responsible for almost 20 percent of all cancer deaths in women. It ranks second in death rate after lung cancer. Roughly 180,000 women are diagnosed with this disease each year, of which 44,000 will die. With increased awareness and increased use of routine mammograms, more women are diagnosed in the earlier stages of this disease, at which time a cure may be possible. First degree relatives of all patients with this cancer should be monitored carefully. This cancer has a tendency to run in families and is associated with genetic abnormalities, for which they can be tested.

Ovarian cancer is much more common in women over the age of 50 years and after menopause. Approximately 25,000 American women develop this kind of cancer and 14,000 die as a result of it every year. Survival of patients depends on the extent of the cancer at the time of initial diagnosis. Most patients with early stages of ovarian cancer can be cured with surgery and chemotherapy. In the majority of patients for whom cure is not possible, survival could vary from months to years, depending on the extent of cancer, the overall condition of the patient, as well as the response to treatments and the duration of the response.

Lung cancer is the second most common malignancy affecting both sexes. Roughly 170,000 Americans are diagnosed with this disease every year. It is considered the most rapidly increasing cause of death from cancer. Since 1987, lung cancer has been the leading cause of cancer death in women, surpassing breast cancer. And while lung cancer incidence has leveled off among men, it continues to rise among women.

Taxol (paclitaxel) is a chemotherapeutic drug used widely for the treatment of breast and ovarian cancer with some potential application for the treatment of lung cancer. But properties of taxol create several problems. The very low water solubility (less than 0.01 mg/ml) leads to decreased absorption of this powerful drug into the bloodstream and additional problems are related to its formulation.
Furthermore, once in the body, it attacks not only the cancer cells, but also normal tissue, such as bone marrow, nerve fibres, and mucous membranes. The degree and

severity of the side effects depend on the dosage and schedule of taxol administration. The following are some of the most common and important side effects: low white blood counts, low platelet count, anemia, hair loss, nerve damage, allergic reaction, and fluid retention. The occurrence of allergic reactions, skin reactions and fluid retention can be significantly reduced by pretreatment of patients with steroids. Since taxol is metabolized in the liver and excreted into bile, the dosage should be substantially reduced in patients with liver dysfunction or massive liver metastasis.

The activities to improve taxol performance increased dramatically over the last ten years. There is a new search for taxol analogues with similar or even improved anticancer activities with less severe side effects and enhanced water solubility (Mathew et al., 1992).

The main objective of this study is applying neural network models to screen a set of structurally similar taxol analogues for their anticancer activity, with the significant prediction of potentially improved performances.

The tested hypothesis is:

Is the antitumor activity of this compound against a particular cell line higher or the same as taxol activity?

The other goal of this study is to confirm the feasibility of this technique to correctly predict new drug candidates as part of the process of 'rational' drug design.

It is often easy to classify compounds according to their activity against particular cell lines once those activities are measured along with other properties of the molecule. However, the neural network approach in this study relies on the calculations of molecular properties so that the potential of a compound can be established even before synthesis is undertaken.

Background

There are three, often-overlapping paths to drug lead discovery: mass screening of products, 'irrational design' procedure using combinatorial chemistry methods, and 'rational' techniques built on quantitative structure-activity relationship (QSAR) design and its variants. Rational drug design is based on the principle that the biological properties of molecules are related to their actual structural features. The understanding of such dependency in terms of QSAR is one of the main goals of medicinal chemistry and plays a key role in modern computer-aided drug design.

Since the introduction of the concept of QSAR in drug discovery in the 1960s (Hansch *et al.*, 1964; Free *et al.*, 1964) and its extensive application to the design of biologically active compounds, the field has developed considerably. New computational methods have been introduced and applied to the design of new molecules and the analysis of chemical and biological data (Kowalski *et al.*, 1979; Wold *et al.*, 1983; Massart *et al.*, 1988; Zupan *et al.*, 1993; Persidis *et al.*, 1997; Ajay *et al.*, 1998; Shi *et al.*, 1998; Tetteh *et al.*, 1999). Molecular modeling has become a new discipline in pharmaceutical research that has contributed to chemically reasonable improvements in the lead compounds, in order to enhance their potency, as well as the discovery of entirely new therapeutic agents.

Quantitative structure-activity relationship (QSAR)

Investigation of the relationship between chemical structure and the activity of compounds helps the understanding of the activity of interest, and may enable a valuable prediction of the activity of new compounds, based on the knowledge of the chemical structure alone. QSAR is used to recognize and utilize the complex connection function between the principal properties of the chemical compounds and their biological, ecotoxicological, or pharmacological activity.

Activity = f (structure)

The principle of QSAR consists in relating the activities observed for a series of chemicals to a set of theoretical parameters, which are assumed to describe the relevant properties of their structure quantitatively. Therefore, the QSAR application requires three classes of fundamental information:

1. Descriptors of the chemical structure.

2. Measurements of the selected activity.

3. Methods to quantify the existing relationship between chemical structures and actual activities.

QSAR describes biological activities in terms of physicochemical and structural characteristics, i.e. steric, electronic, and hydrophobic properties of the molecules within a congeneric data set in the following fashion:

$$\text{Bioactivity} = f\{\Sigma(\text{steric}) + \Sigma(\text{electronic}) + \Sigma(\text{hydrophobic})\} \text{ interactions.}$$

There are two basic approaches in QSAR modeling techniques.

1. **Hansch analysis.** The QSAR investigation is performed for the whole molecule based on physicochemical descriptors of the molecular properties, such as molecular weight, molar refractivity, partition coefficient (LogP), steric effects, and molecular topology (Kubinyi, 1993). The analysis is based on the assumption that similar physicochemical properties are expected to also have similar activities, and different molecular substituents change the properties of the whole molecule.

2. **Free-Wilson model.** This approach for QSAR is based on the hypothesis that the activity for a set of similar molecules can be described by additive properties of the activity contributions from substituents or structural elements, present in a parent structure (Free *et al.*, 1964). Thus, the presence or the absence of structural elements is indicated by the values 1 and 0, respectively. The model is suitable for analysis of chiral compounds and

chemical isomers. However, the Free-Wilson technique has some severe limitations. The method is strictly interpolative from a structural point of view; i.e. only structural properties described in the model by the set of chemical compounds can be used. The method cannot, as opposed to techniques based on a physicochemical description of the structures, be used for any extrapolative purposes to find new structural elements that are predicted to be of potential interest. It is questionable especially in case of a single-point determination (a particular structure feature that occurs only once in the data set), because the corresponding structural contribution will then contain the entire experimental error of the dependent variable, e.g. biological activity.

For the QSAR model, based on the pattern recognition neural network architecture, the Hansch analysis is used as all the descriptors were calculated for each taxol analogue molecule.

Taxol analogues

Input data for the neural network design

Structures and anticancer activities for all compounds used in a design of the optimal neural network prototype were obtained from National Cancer Institute (NCI), Rockville, Bethesda, MD. A search of the July 1998 database containing 29,969 compounds tested *in vitro* in the anticancer screen, can be directed for a set of compounds grouped by mechanism of action, similar chemical structure, or NSC recognition number (NCI internal identification number). The NCI *in vitro* primary cancer screen consists of a panel of 60 different human tumor cell lines of 9 types of cancer (leukemia, non-small cell lung, colon, central nervous system, melanoma, ovarian, renal, prostate, and breast cancer) against which compounds are tested over a defined range of concentrations to determine the relative degree of growth inhibition or cytotoxicity against each cell line.

For the neural network design, 50 taxol analogues were chosen, based on molecular similarity, i.e. each compound has the basic structure of baccatin III (Figure 3A), characterized by its taxane ring system with a four-membered oxetane ring and ester side chain at the position C-13.

Taxol analogues (61 compounds) were obtained from Prof. G. Georg, Department of Medicinal Chemistry, The University of Kansas, Lawrence, 66245 to test the prototype. These compounds are structurally related to the taxol molecule, with different substituent groups in position C-7 or C-10.

Output data – anticancer activities

Three different types of cancer: ovarian (OVA), breast (BRE), lung non-small cell (LNS) cancer, and an average across all 60 cell lines, represented by the index GI_{50} values, are the prime targets of the QSAR and *in vitro* testing procedure. The central point is to reveal QSAR between the molecular structure of a compound and its anticancer activity towards a specific type of cancer, using pattern recognition back-propagation neural network (BPNN) prototype, and to generate significant prediction of anticancer activities for all tested compounds.

Assay for determination of anticancer activity

In all *in vitro* assays, exposure to an antitumor compound decreases the number of potential tumor cells by direct cell killing, or by solely decreasing the rate of cellular proliferation. The tumor cells are inoculated over a series of standard 96-well microtiter plates on day zero, with relatively low initial cell inoculation densities (typically 5000 – 40,000 cells per well based on growth parameters). The cells are incubated on the microtiter plate for 24 hours at 37°C for stabilization. (Monks *et al.*, 1991). The tested compounds are then added to the wells in five 10-fold dilutions starting with the highest soluble concentration (usually 10^{-4} to 10^{-8} M). The assay involves an incubation of either the chemical agents or extracts for 48 hours in 5% CO_2 atmosphere at 100% humidity with the tumor cell lines. At the termination point, the cells are assayed by the sulforhodamine B procedure (Boyd *et al.*, 1995). The cells are washed and the remaining dye is a function of the adherent cell mass.

Optical densities are measured on an automated plate reader. The data are then analyzed into special concentration parameters: GI_{50}, TGI, and LC_{50}.

The 50% growth inhibition parameter (GI_{50}) is the concentration of the test agent calculated by the equation

$$100 \times (T-T_0)/(C-T_0) = 50\% = \textit{percentage growth (PG)}$$

where T is the optical density of the test well after the 48-hour period of drug exposure, T_0 represents the optical density at initial time zero, and the control optical density is C. GI_{50} is usually called a growth-inhibitory power of the test agent.

The TGI indicates a 'total growth inhibition' or cytostatic level of effect. It is a drug concentration where T and T_0 are the same, so the fraction (*percentage growth*) is equal zero.

The LC_{50} is the lethal concentration, 'net cell killing' or cytotoxicity parameter. It is a concentration where

$$100 \times (T-T_0)/T_0 = -50\% = PG$$

The control optical density is not used in this calculation.

These concentration values are calculated by interpolation on the concentration axis using the tested concentrations that give PG values above and below the reference values (e.g. 50 for GI_{50}).

Currently, about 45% of the GI_{50} records in the NCI database are approximated in this way (http://dtp.nci.nih.gov/). If, however, for a given cell line, all of the tested concentrations indicate percentage growths above the reference level (+50, 0, -50), then the lowest tested concentration (specified in negative $\log_{10}$ units) is automatically assigned as the default value.

Neural network prototype

Physicochemical descriptors

The biological activity can be regarded as a complex function of the physicochemical and structural properties of a ligand molecule. The structure of the compound is the only source of input data. The initial set of 27 features for each compound were calculated using CAChe software, version 3.1 (Computer-*aided* *chem*istry, Oxford Molecular Group, Inc. 2105 South Bascom Ave., Campbell, CA 95008; http://www.oxmol.com). The experiments in CAChe use mathematical models, derived from computational chemistry, to calculate molecular physicochemical properties and topological indexes. The computation chemistry tools are acquired from classical mechanics and quantum mechanics, and are applied in the same manner to the training and testing data sets. All calculated physicochemical properties (27 features) are summarized in Table 1, Chapter 3.

Reducing the pool of 27 original variables to an appropriate size eliminates those descriptors which contribute either no (or very limited) information or whose information is redundant with that of other characteristics present in the final pool. Any variable that had identical or zero value for more then 90% of the compounds could be practically eliminated (Blankley, 1996). Based on the results of each performed data analysis (correlation matrix, principal component, and pattern analysis), the initial dimension of input data set with all variables was efficiently reduced to a final set of 9 descriptors, without any serious loss of information.

A molecular orbital calculation yields a set of eigenvalues or energy levels, in which all the available electrons are accommodated. The highest filled energy level is called the HOMO (*h*ighest *o*ccupied *m*olecular *o*rbital). The second highest energy level, which is unoccupied because no more electrons are available, is the LUMO or *l*owest *u*noccupied *m*olecular *o*rbital. Both energy parameters have been removed from the initial pool of characteristics since they exhibit relatively small pattern variability, as all the data remain within a relatively short interval. The topological characteristics (molecular weight, atom and bond counts, connectivity and shape indexes) are also excluded from the final set of variables, since they are highly dependent and reveal an almost identical pattern.

Table 1 displays the final set of physicochemical properties used in the process of designing the neural network prototype for screening anticancer activities in taxol analogues.

Physicochemical property	Abbreviation
Lipophilicity	LogP
Molar refractivity	MR
Dipole moment	DM
Dipole vector X	DV-X
Dipole vector Y	DV-Y
Dipole vector Z	DV-Z
Steric energy	SE
Conformation minimum energy	CME
Heat of formation	HOF

Table 1 Physicochemical parameters used for the design of the neural network prototype.

Lipophilicity (hydrophobicity, partition coefficient, LogP) is an important molecular property, which is related to the ability of a compound to partition between water and nonpolar solvent. The relative affinity of a drug molecule for an aqueous or lipid medium is an important correlate of drug activity due to the absorption, transport, and partitioning phenomena. LogP, the logarithm of the partition coefficient between water and 1-octanol, has been used to define lipophilicity (Hansch *et al.*, 1979)

$$\text{LogP} = \log\{[\text{drug}]_{\text{octanol}} / [\text{drug}]_{\text{water}}\}$$

where in this model $[\text{drug}]_{\text{octanol}}$ is the concentration of a solute in the lipid phase approximated by 1-octanol, and $[\text{drug}]_{\text{water}}$ is the concentration of the solute in the aqueous phase. The lipophilicity of a drug affects a number of pharmacokinetic parameters, since it normally has to pass various biological membranes by passive diffusion. A more lipophilic drug is better adsorbed after oral administration than a less lipophilic analogue. Similarly, lipophilic drugs more readily pass the blood-brain barrier (BBB) and are therefore better distributed to the brain. An optimal lipophilicity for penetration through the BBB appears to be about 2. This characteristic has been shown to be highly correlated with a diversity of biological activities, including antitumor activity and toxicity (Grunenberg *et al.*,1995; Domine *et al.*, 1996). It is evident that LogP (usual values from -4 to 8) plays a vital role in the interaction between drugs and their receptors. The lipophilic quality of drugs is also an important factor in drug metabolism. In addition, the absorption, and excretion of many classes of drugs are highly dependent on the lipophilicity coefficient.

Molar refractivity (MR) can be viewed as a steric descriptor of a molecule. Molecular polarizability and molar refractivity are closely related properties that are a measure of a molecule's susceptibility to becoming polarized. MR can be calculated from the refractive index and the molar volume according to the Lorentz-Lorentz equation:

$$MR = [(n^2-1)/(n^2+2)]\ (MW/d)$$

In this expression, n represents the refractive index, MW is the molecular weight, and d is the density of a compound, while (MW/d) represents the molar volume. However, applications in QSAR usually employ empirical estimates (Miller, 1990), based on atomic, bond, or group contributions. The larger the polar part of a molecule, the larger its MR value.

Electric dipole moment (DP) encodes the strength of polar type interactions. Its value is estimated in CAChe software by quantum mechanical techniques. Dipole vectors then characterize the geometrical arrangement of a compound in three-dimensional space.

Steric energy (SE) of a molecule represents the sum of the molecular mechanics potential energies calculated for bonds, bond angles, and so forth. It is specific to the Mechanics part of CAChe and depends on the force-field used.

Conformation minimum energy (CME) is calculated from the optimized conformation of the chemical sample. Depending on which procedure is used, the calculated energy may be steric energy, heat of formation, or total energy. The optimization procedures usually locate a minimum energy conformation near the starting (optimal) geometry.

Heat of formation (HOF) denotes the energy released or used when a molecule is formed from elements in their standard states. The energy variables are calculated in units of kcal/mol.

Figure 1 shows a profile of the training set of 50 compounds, where the scaled values of descriptors are plotted against the final set of structural properties in the case of ovarian cancer. The profile reveals the whole complexity of the QSAR problem with heavy overlapping regions. The compounds are presented in a simplified way as binary data, where class 1 represents the analogues with the same or higher activity than taxol, while class 0 designates the subset of compounds with a lower activity than taxol. As the profile shows, it is very difficult to distinguish between both classes.

Figure 2 illustrates the profile of averages calculated for both classes against each physicochemical characteristic (in ovarian cancer) for all trained compounds with the same class identification. The overlap occurs in two regions (variables MR and DV-Z) and could justify the limits of the multivariate analysis applied in this model.

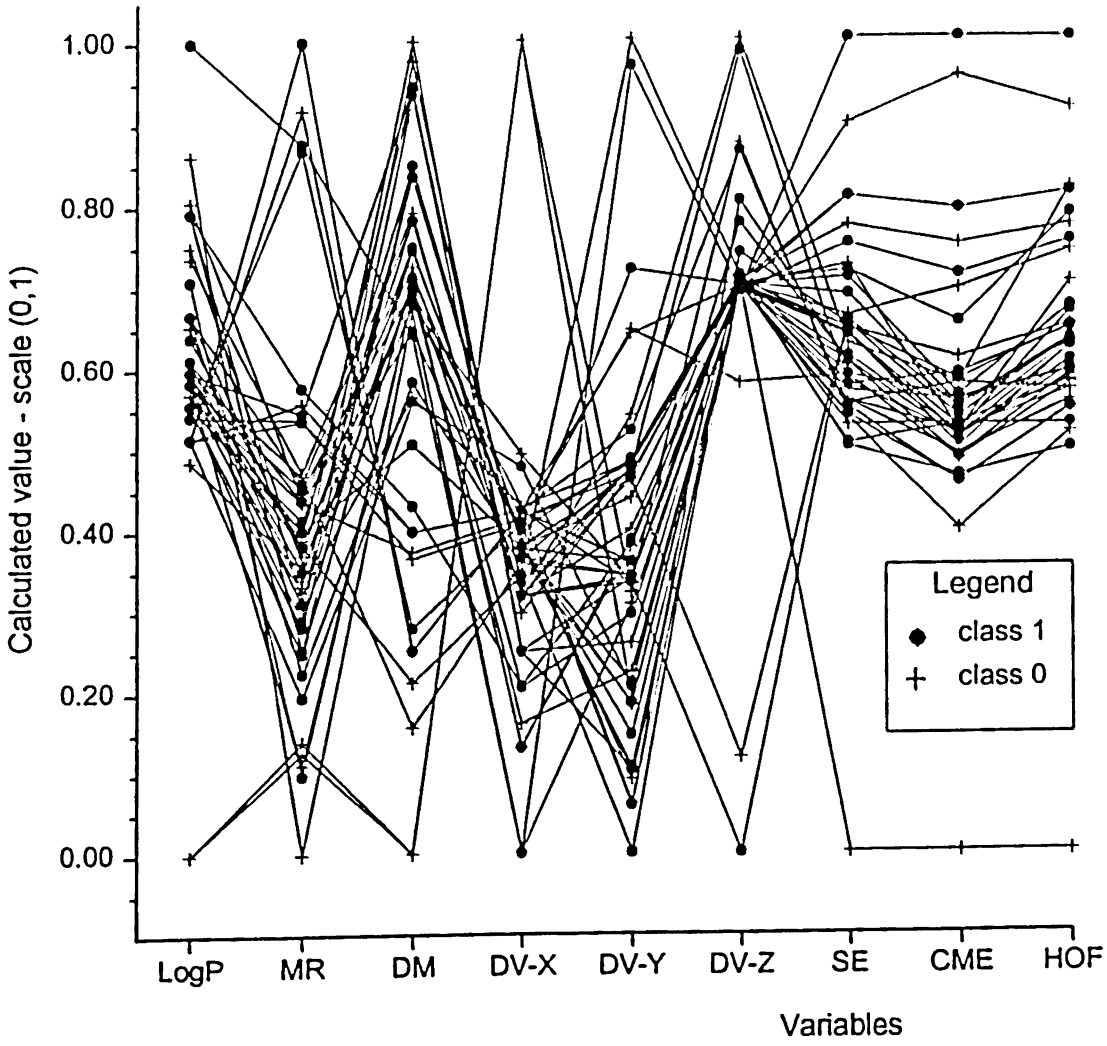


Figure 1. Profile of the training data set (OVA)

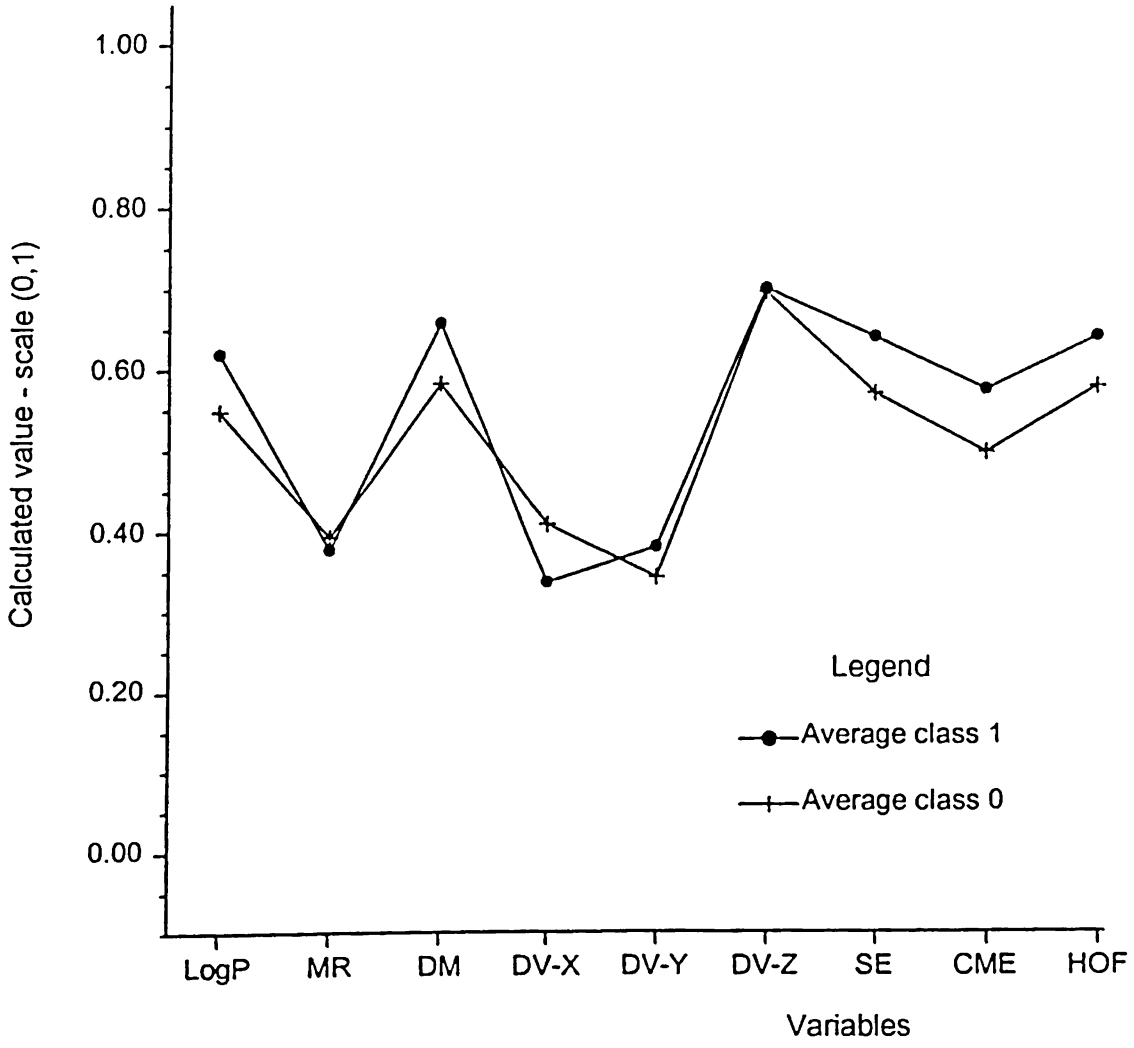


Figure 2. Profile of the averages for both classes (OVA)

Neural network architecture and accuracy

The training set consists of randomly selected 40 chemical compounds (feature vectors) with 9 physicochemical descriptors. Ten compounds are in the testing set to evaluate the predictive power and the accuracy of the neural network (NN) model.

The NN is simulated, trained, and tested using MATLAB software (version 5.3.0, 1999). Feed-forward back propagation neural network (BPNN), applying the gradient descent method with one hidden layer (and selected number of units), together with one output unit (antitumor activity), is used.

The prediction of the NN prototype is considered incorrect when the predicted value differs from the actual value more than ±0.1. The systematic approach of selecting the NN variables, followed by the experimental optimization (i.e. longer training in order to improve the NN performance), results in an accuracy of 92% against ovarian and lung cancer, 94% for breast cancer, and 90% for the index GI_{50}.

In terms of correlation between predicted and actual outputs, BPNN accuracy is 0.831 for ovarian, 0.945 for lung, and 0.913 for breast cancer, while the index GI_{50} exhibits an accuracy of 0.886. A comparative molecular field analysis (CoMFA) was performed on the subset of taxol analogues (Czaplinski *et al.*, 1994) with the accuracy correlation of 0.853. By comparison of accuracy correlations of both procedures, the results clearly indicate, that BPNN has in most cases higher predictive accuracy than the CoMFA method.

Testing the neural network prototype

Input data

Taxol analogues (total of 61 compounds, where 13 analogues have different substituent groups on position C-7 of the taxol molecule, while 48 analogues represent the group of substituents on position C-10) are treated in the same way as raw data for the NN prototype. All physicochemical features for each compound were calculated using CAChe software. The data were then scaled according to the training set because the tested vectors should be consistent with the initial input vectors. Original set of tested compounds was at that point divided into two parts.

Thirty-five tested compounds are within the training <0,1> continuous interval, while 26 analogues exceed this interval in both ways, i.e. negative value or a value greater than 1. Since the extrapolation over the reasonable limit (i.e. in all variables) could cause deficient generalization, the anticancer activity prediction was performed only for the analogues consistent with the initial training data set. Testing data were transferred to the trained BPNN prototype and the output (anticancer activity) predictions were reported.

Anticancer activity prediction

Table 2 shows predicted activities of tested analogues (i.e. within the region of the trained BPNN) against ovarian cancer. Bolded rows express the measured activities for taxol and taxotere against a particular cancer type.

Analogue	Predicted activity
10y110905	0.7894
10y110963	0.7861
10y110938	0.7827
10y110964	0.7810
10y001127	0.7800
10y110913	0.7785
10y110901	0.7732
10y110968	0.7655
07y001119	0.7550
10y110939	0.7511
10y110949	0.7400
10y110917	0.7377
10y110918	0.7261
10y110931	0.7190
07y122601	0.7120
10y110920	0.7119
10y110954	0.7061
10y110941	0.6959
10y110961	0.6943
10y110955	0.6891
10y110951	0.6846
10y110936	0.6833
10y110947	0.6812
10y110921	0.6514
10y110919	0.6506
10y110942	0.6397
Taxotere	**0.6311**
Taxol	**0.6161**

Table 2 Predicted activity against ovarian cancer

By comparison, most of the tested analogues (26 out of 35) suggest predicted activity against ovarian cancer higher than taxol and taxotere.

The NN predictions of analogues tested against lung cancer are shown in Table 3.

Analogue	Predicted activity
10y110938	0.8947
10y110905	0.8912
10y110964	0.8754
07y001119	0.8705
10y110913	0.8671
10y110963	0.8648
10y001127	0.8590
10y110901	0.8409
10y110918	0.8389
07y122601	0.7209
Taxotere	**0.6817**
10y110939	0.6640
Taxol	**0.6547**

Table 3 Predicted activity against lung cancer

Table 4 represents the NN predictions against breast cancer.

Analogue	Predicted activity
10y110938	0.8083
Taxotere	**0.7289**
10y110943	0.7280
10y110939	0.6524
10y110937	0.6401
Taxol	**0.5878**

Table 4 Predicted activity against breast cancer

Predictions of overall average, i.e. index GI_{50}, are illustrated in Table 5.

Analogue	Predicted activity
10y110905	0.6827
10y110963	0.6825
10y001127	0.6795
10y110913	0.6765
10y110964	0.6757
10y110918	0.6754
10y110901	0.6743
10y110941	0.6729
10y110968	0.6670
10y110949	0.6661
10y110947	0.6628
10y110931	0.6617
10y110954	0.6596
10y110951	0.6595
10y110939	0.6587
10y110917	0.6374
10y110961	0.6371
Taxotere	**0.6064**
Taxol	**0.6013**

Table 5 Predicted activity against the average index GI_{50}

Since the BPNN model is trained on taxol analogues obtained from NCI, the final step is to combine all the results together in tabulated form, to show which of the tested analogs has the highest activity. The next two tables show twenty compounds sorted by their activities, which are predicted (in case of compounds labeled by italic character) and measured (in case of NCI compounds). Again, taxol and taxotere are included for a comparison.

OVA Analogue	OVA Activity	LNS Analogue	LNS Activity
671864	0.8569	671864	0.9067
673185	0.8096	*10y110938*	0.8947
10y110905	0.7894	*10y110905*	0.8912
10y110963	0.7861	*10y110964*	0.8754
10y110938	0.7827	*07y001119*	0.8705
10y110964	0.7810	*10y110913*	0.8671
10y001127	0.7800	*10y110963*	0.8648
10y110913	0.7785	*10y001127*	0.8590
10y110901	0.7732	*10y110901*	0.8409
662160	0.7710	*10y110968*	0.8389
10y110968	0.7655	673185	0.8172
647753	0.7638	662160	0.8120
07y001119	0.7550	647753	0.7545
10y110939	0.7511	674282	0.7377
10y110949	0.7400	664403	0.7351
10y110917	0.7377	*07y122601*	0.7209
651195	0.7377	651195	0.7170
10y110918	0.7261	673191	0.7057
10y110931	0.7190	666607	0.7054
07y122601	0.7120	673193	0.6928
Taxotere	**0.6311**	**Taxotere**	**0.6817**
Taxol	**0.6161**	**Taxol**	**0.6547**

Table 6 Analogues with high activities against OVA and LNS

Table 7 presents the taxol analogues, together with taxol and taxotere, with predicted and measured (NCI compounds) antitumor activities against breast cancer (BRE) and average overall index GI_{50}.

BRE Analogue	BRE Activity	GI_{50} Analogue	GI_{50} Activity
10y110938	0.8083	647753	0.7670
Taxotere	**0.7289**	662160	0.7514
10y110943	0.7280	673185	0.7348
674282	0.7187	*10y110905*	0.6827
671864	0.7175	*10y110963*	0.6825
662160	0.7055	*10y001127*	0.6795
666607	0.6931	671864	0.6784
10y110939	0.6524	651195	0.6776
10y110937	0.6401	*10y110913*	0.6765
673185	0.6363	*10y110964*	0.6757
671869	0.6099	*10y110918*	0.6754
662161	0.6071	*10y110901*	0.6743
673191	0.6042	*10y110941*	0.6729
671867	0.5960	664403	0.6681
671872	0.5912	664401	0.6681
671871	0.5880	*10y110968*	0.6670
Taxol	**0.5878**	*10y110949*	0.6661
687962	0.5816	*10y110947*	0.6628
671868	0.5637	*10y110931*	0.6617
673193	0.5634	*10y110954*	0.6596
689292	0.5550	**Taxotere**	**0.6064**
608832	0.5501	**Taxol**	**0.6013**

Table 7 Analogues with high activities against BRE and GI_{50}

In the case of breast cancer, one analogue (10y110938) reveals even higher predicted activity than taxotere and other NCI compounds, while compound (10y110943) shows an almost similar activity value.

Results and discussion

The results, in terms of predicted antitumor activities, are based on the generalization capability of trained neural network optimal prototype. Predictions are made within a selected error ($\Delta = \pm 0.1$), where the error criterion was chosen as a part of the BPNN architecture design process. Table 8 reveals ten compounds (with the highest predicted activities) for each type of cancer.

Order	OVA	LNS	BRE	GI_{50}
1.	10y110905	10y110938	10y110938	10y110905
2.	10y110963	10y110905	Taxotere	10y110963
3.	10y110938	10y110964	10y110943	10y001127
4.	10y110964	07y001119	10y110939	10y110913
5.	10y001127	10y110913	10y110937	10y110964
6.	10y110913	10y110963	Taxol	10y110918
7.	10y110901	10y001127	10y110913	10y110901
8.	10y110968	10y110901	10y110964	10y110941
9.	07y001119	10y110968	10y110905	10y110968
10.	10y110939	07y122601	10y001127	10y110949

Table 8 Ordered taxol analogues with the highest predicted activities for each cancer type

Figure 3 shows the structural formulas of baccatin III, taxol, and taxotere, while Figure 4 illustrates the position for the substituents in taxol molecule for 7-O and 10-O analogues.

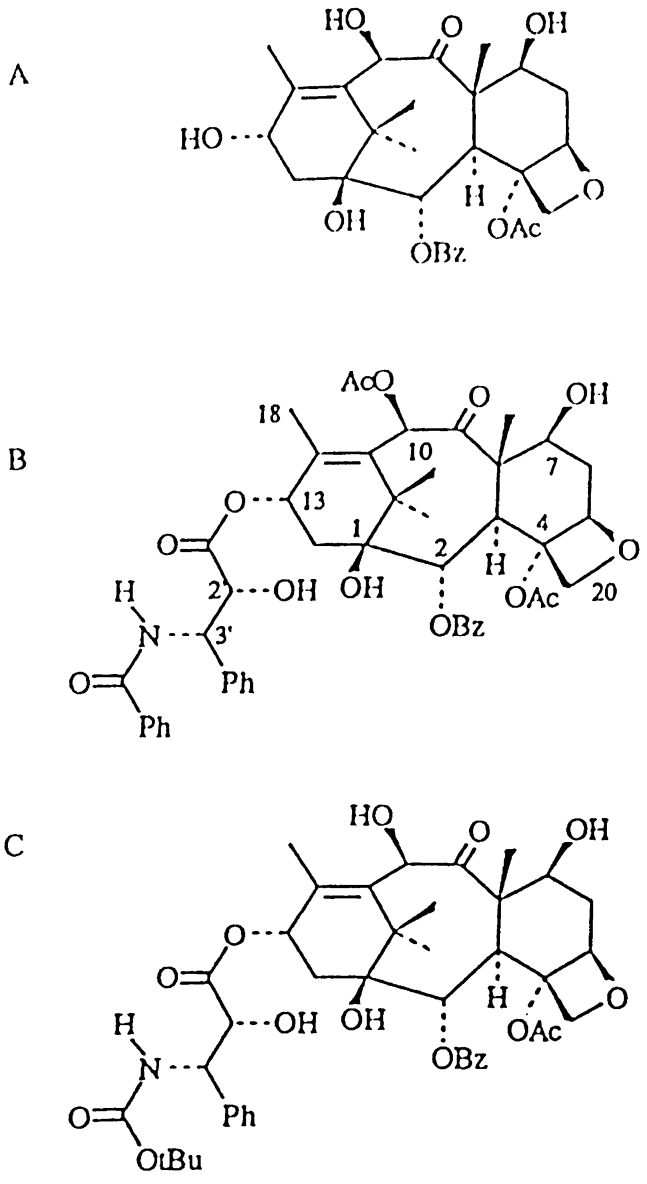


Figure 3. Structural formulas of the baccatin III (A), taxol (B), and taxotere (C) molecules

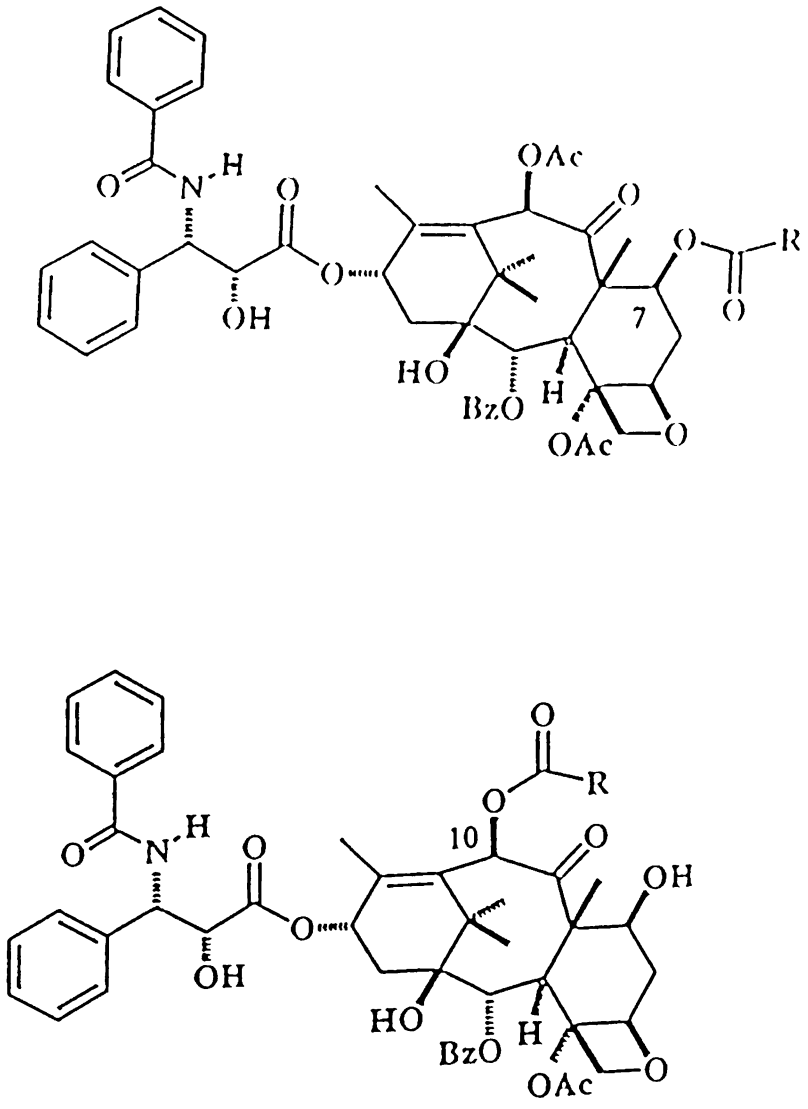


Figure 4. Structures of the 7-O and 10-O taxol analogues

Table 9 shows highly active taxol analogues sorted by the number of compound and the adequate substituent in position C-10 (10-O-analogues) or C-7 (7-O-analogues). This tabulated arrangement may be useful in answering fundamental questions, such as, what are these compounds and what do they have in common?

Analogue	Substituent
10y110905	propionic acid
10y110913	Cyclopropanecarboxylic acid
10y110937	pyrrole-2-carboxylic acid
10y110938	1-methyl-2-pyrrolecarboxylic acid
10y110939	2-furoic acid
10y110943	3-thiophencarboxylic acid
10y110963	crotonic acid
10y110964	acylic acid
10y001127	Ethylformate
07y001119	4-aminobenzoic acid

Table 9 Taxol analogues with the highest predicted activities

Baccatin III is natural taxane, isolated from the roots of *Taxus baccata* in 1965. This taxol analogue (without the ester side chain at C-13) is substantially less active (activity less than 0.1 on the (0,1) scale) than taxol and taxotere (activities around 0.6), indicating the importance of the side chain for antitumor activity (Wani *et al.*, 1971; Georg *et al.*, 1992a; Kingston, 1994).

Structural modifications along the upper part of the taxol molecule (C-6 to C-12) do not seriously affect the taxol bioactivity, assuming that this region is not closely involved in binding to tubulin.

The 7-hydroxyl group can be modified easily and selectively without serious loss of biological activity. Analogues with substituents containing an ionizable group (amines, phosphates, carboxylic, sulfonic and amino acids) maintain biological activity (Lataste *et al.*, 1984; Mathew *et al.*, 1992). They were used in an effort to develop active analogues with greater water solubility (Deutsch *et al.*, 1989; Larsen, 1997).

The high antitumor activity predicted for the analogue with a 4-aminobenzoic acid substituent in C-7 position is in good agreement with the reported high activity in microtubule assays for benzoic acid (Kingston *et al.*, 1990) and p-azidobezoic acid (Georg *et al.*, 1992b).

The acetyl group at the C-10 hydroxyl group is not essential for antitumor activity as highly active taxotere lacks the C-10 acetyl group. The tolerance to C-10 structural variations is demonstrated by comparison of 10-deacetyltaxol, taxotere (both with a hydroxyl group in C-10), and 10-acetyltaxotere (OAc in C-10). These three analogues maintain very similar activity in the microtubule assay (G. Georg *et al.*, 1995).

Table 10 presents the structures of the substituent groups for the taxol analogues with the highest predicted activities.

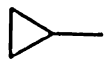
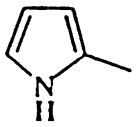
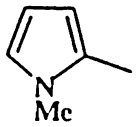
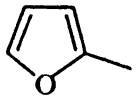
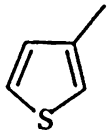
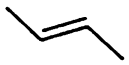
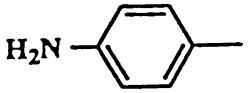
Analogue	Substituent	Structure
10y110905	propionic acid	CH_3CH_2
10y110913	cyclopropanecarboxylic acid	
10y110937	pyrrole-2-carboxylic acid	
10y110938	1-methyl-2-pyrrolecarboxylic acid	
10y110939	2-furoic acid	
10y110943	3-thiophencarboxylic acid	
10y110963	crotonic acid	
10y110964	acylic acid	
10y001127	ethylformate	EtO
07y001119	4-aminobenzoic acid	

Table 10 The structures of the substituent groups

Compounds with the highest predicted antitumor activities, 10y110905, 10y110938, and 10y110963 are analogues with the following substituents: propionic, 1-methyl-2-pyrrolecarboxylic, and crotonic acid in position C-10 (10-O-analogues). This indicates that short alkane and alkene chains, a cycloalkane ring, in the case of cyclopropane, or a heterocyclic pyrrole ring and its derivatives located in C-10-O position could be the best choice for future highly potent antitumor taxol analogues. On the other hand, the compounds with benzoic acid and its derivatives or a long alkane chain, (octanoic acid) reveal the lowest predicted antitumor activities.

The oxetane ring is relatively inert chemically, and it has been suggested (Suffness, 1993) that its role simply might be to act as a lock and to preserve the conformation of the diterpenoid ring system of taxol.

Conclusion

Testing of taxol analogues starts with the calculations of the physicochemical structural parameters, which are related to the antitumor activities throughout the QSAR (activation, transfer function). Since the structural properties of the tested compounds are unknown, the best estimations are the only calculated values. Introducing the measured assessments of compounds into the training or the testing data set will lead in increasing error based on the differences between the measured and calculated properties of molecules and additional complications. But the most important reason for not using measured property values is that this type of analysis has a very significant prediction power so the potential of the compound activities can be established prior to the synthesis.

The process continues with the transformation of the input 'raw' data (using min/max procedure) into the continuous <0,1> interval the same way as for a training data set. The network system should not be allowed to extrapolate in all dimensions (variables), i.e. it should not be used to predict activities of input vectors, which are outside the scope of the training input space. It is important to avoid this type of error and test only the compounds within the training interval for most physicochemical properties. In other words, the outliers from the testing data set are not used for the assessment of the anticancer activity, since the power of the prediction could be substantially decreased.

The test procedures are based on the neural network prototypes, which are properly trained on 50 NCI taxol analogues. Tested compounds are then sorted by the predicted anticancer activities and presented in tabulated forms with the combination of network data (i.e. NCI compounds), to determine the analogues with the highest activities against a particular kind of cancer and the index GI_{50}.

Compounds with the highest predicted antitumor activities, 10y110905, 10y110938, and 10y110963, are taxol analogues with a substituent propionic, 1-methyl-2-pyrrolecarboxylic, and crotonic acid in position C-10 (10-O-analogues). These compounds should be synthesized following testing, where the actual antitumor activity should be measured, in order to verify the prediction power of the neural network prototype.

References

Ajay; Walters, W.P.; Murcko, M. Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? *J. Med. Chem.*, **1998**, *41*, 3314-3324.

Blankley, C.J. Recent development in 3D-QSAR. In. *Structure-property correlations in drug research* (H. van de Waterbeemd, Ed.). R.G. Landes Co., Austin, TX, **1996**, Chapter 5, 111-177.

Boyd, M.R.; Paull, K.D. Some practical considerations and applications of the National Cancer Institute *in vitro* anticancer drug discovery screen. *Drug Development Research*, **1995**, *34*, 91-109.

Czaplinski, K.H.; Grunwald, G.L. A comparative molecular field analysis (CoMFA) derived model of the binding of Taxol analogues to microtubules. *Bioorganic & Medicinal Chemistry Letters*, **1994**, *4 (18)*, 2211-22216.

Deutsch, H.M.; Glinski, J.A.; Hermandez, M.; Haugwitz, R.D.; Narayanan, V.L.; Suffness, M.; Zalkow, L.H. Synthesis of congeners and prodrugs. III. Water-soluble prodrugs of Taxol with potent antitumor activity. *J. Med. Chem.*, **1989**, *32*, 788-802.

Domine, D.; Wienke, D.; Devillers, J.; Buydens, L. A new nonlinear neural mapping techniques for visual exploration of QSAR data. In, *Neural Networks in QSAR and Drug Design* (J. Devillers, Ed.). Academic Press, London, **1996**, 223-253.

Free, S.M.; Wilson, J.W. A mathematical contribution to structure – activity studies. *J. Med. Chem.*, **1964**, *7(4)*, 395-399.

Georg, G.I.; Cheruvallath, Z.S.; Himes, R.H.; Mejillano, M.R.; Burke, C.T. Synthesis of biologically active taxol analogues with modified phenylisoserine side chains. *J. Med. Chem.*, **1992a**, *35*, 4230-4237.

Georg, G.I.; Harriman, G.C.B.; Himes, R.H.; Mejillano, M.R. Taxol photoaffinity label: 7-(p-azidobenzoyl)taxol. Synthesis and biological evaluation. *Bioorg. Med. Chem. Lett.*, **1992b**, *2*, 735-738.

Georg, G.I.; Boge T.C.; Cheruvallath, Z.S.; Clowers, J.S.; Harriman, G.C.B.; Hepperle, M.; Park, H. The medicinal chemistry of Taxol. In, *Taxol: Science and Application.* (M. Suffness, Ed.), CRC Press, Inc., Boca Raton, Florida 33431, **1995**, 317-379.

Grunenberg, J.; Herges, R. Prediction of chromatographic retention values (R_M) and partition coefficients (LogP) using a combination of semiempirical self-consistent reaction field calculations and neural networks. *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 905-911.

Hansch, C.; Fujita, T. Pi-sigma-rho analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **1964**, *86*, 1616-1626.

Hansch, C.; Leo, A.J. Substituent constants for correlation analysis in chemistry and biology, John Wiley & Sons, New York, **1979.**

Kingston, D.G.I.; Samaranayake, G.; Ivey, C.A. The chemistry of Taxol, a clinically useful anticancer agent. *J. Nat. Prod.*, **1990**, *53*, 1-44.

Kingston, D.G.I. Taxol: The chemistry and structure-activity relationships of a novel anticancer agent. *TIBTECH*, **1994**, *12*, 222-227.

Kowalski, B.R.; Gerlach, R.W.; Wold, H. Partial least-squares path modeling with latent variables. *Anal. Chim. Acta*, **1979**, *112*, 417-421.

Kubinyi, H. QSAR: Hansch analysis and related approaches. In, *Methods and principles in medicinal chemistry.* (R. Mannhold, P. Krogsgaard-Larsen, and H. Timmermann Eds.), VCH Publishers, New York, **1993.**

Larsen, I.K. Anticancer agents. In, *A textbook of drug design and development.* (P. Krogsgaard-Larsen, T. Liljefors, and U. Madsen Eds.), Harwood Academic Publishers, Amsterdam, **1997,** 460-508.

Lataste, H.; Senilh, V.; Wright, M.; Guenard, D.; Potier, P. Relationship between the structures of Taxol and Baccatin III derivatives and their *in vitro* action on the disassembly of mammalian brain and *Physarum* amoebal microtubules. *Proc. Nat. Acad. Sci.*, **1984,** *81,* 4090-4112.

Massart, D.L.; Vandeginste, B.G.; Deming, S.N. Chemometrics: a Textbook. Elsevier, Amsterdam, **1988.**

Mathew A.E.; Mejillano, M.R.; Nath, J.P.; Himes, R.H.; Stella, V.J. Synthesis and evaluation of some water-soluble prodrugs and derivatives of taxol with antitumor activity. *J. Med. Chem.*, **1992,** *35,* 145-151.

Miller, K.J. Techniques to estimate molecular polarizabilities in QSAR. *J. Am. Chem. Soc.*, **1990,** *112,* 8533-8542.

Monks, A.; Scudiero, D.; Skehan, P.; Schoemaker, R.H.; Paull, K.D.; Vistica, D.; Hose, C.; Langley, J.; Cronise, P.; Vaigro-Wolfe, A; Gray-Goodrich, M.; Campbell, H.; Boyd, M.R. Feasibility of high-flux anticancer drug screen utilizing a diverse panel of human tumor cell lines in culture. *J. Natl. Cancer Ins.*, **1991**, *83*, 757-766.

Persidis, A.; Persidis.A. Artificial intelligence for the drug design. *Nature Biotechnology*, **1997**, *15*, 1035-1036.

Shi, L.M.; Fan, Y.; Myers, T.G.; O'Connor, P.M.; Paull, K.D.; Friend, S.H.; Weinstein, J.N. Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer Ellipticine Analogues. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 189-199.

Suffness, M. Taxol: from discovery to therapeutic use. *Annu. Rep. Med. Chem.*, **1993**, *34*, 304-326.

Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. Quantitative structure-property relationships using radial basis function neural network. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 491-507.

Wani, M.C.; Taylor, H.L.; Wall, M.E.; Coggon, P.; McPhail, A.T. Plant antitumor agents. VI. The isolation and structure of taxol, a novel antileukemic and antitumor agent from *Taxus brevifolia*. *J. Am. Chem. Soc.*, **1971**, *93*, 2325-2327.

Wold, S.; Albano, C.; Dunn, W.J. Pattern recognition: finding and using patterns in multivariate data. In, *Food research and data analysis*. (H. Martens and J.H. Russwurm, Eds.), Applied Science Publication, London, **1983**, 147-188.

Zupan, J.; Gasteiger, J. Neural networks for chemists – an introduction. VCH Publishers, New York, **1993**.