2009

Design of Novel Anticancer Drugs Utilizing Busulfan for Optimizing Pharmacological Properties and Pattern Recognition Techniques for Elucidation of Clinical Efficacy

Ronald Bartzatt

In: Anticancer Drugs: Design, Delivery and Pharmacology ISBN 978-1-60741-004-1
Editors: Peter Spencer and Walter Holt, pp.

Chapter VI

Design of Novel Anticancer Drugs Utilizing Busulfan for Optimizing Pharmacological Properties and Pattern Recognition Techniques for Elucidation of Clinical Efficacy

Ronald Bartzatt[1]
University of Nebraska, College of Arts and Sciences,
Chemistry Department, USA

Abstract

Chronic myelogenous leukemia (CML) is a disorder in which an excessive number of blood stem cells develop into the white blood cell group called granulocytes. The anticancer drug Busulfan is a cell cycle non-specific alkylating agent which is utilized to maintain white blood cell counts below 15000 cells/microliter. The side effects induced by busulfan are significant and affirms the intimation for new drug constructs. Fifteen analogous compounds were generated from the molecular structure of busulfan . These compounds retain the double methanesulfonate functional groups descriptive of this class of alkylating anticancer drugs. However, the carbon chain substituent separating the methanesulfonate is highly modified in order to allow significant changes in drug properties that may produce favorable characteristics that benefit clinical application. Important properties such as Log P, polar surface area, formula weight, molecular volume, Log BB, and violations of the Rule of 5 were determined to ascertain similarities and differences. All fifteen analog compounds retained zero violations of the Rule of 5, which suggests favorable properties for useful drug availability. Values of Log BB and BB remained the same throughout at -0.841 and 0.144, respectively. In addition, values of polar surface area and number of oxygens and nitrogens remained the same throughout at 86.752 A^3 and 6, respectively. However, formula weight, number of atoms, number of rotatable bonds varied significantly with Log P varying across a broad range (-0.428 to

[1] 6001 Dodge Street, Omaha NE 68182 USA. FAX: 402-554-3888. E-mail: rbartzatt@mail.unomaha.edu.

2.734). The variance in Log P within this group of methane sulfonate compounds allows new and potentially highly beneficial pharmacological properties for clinical application. Pattern recognition techniques such as cluster analysis, non-metric multidimensional scaling, discriminant analysis, and K-means cluster analysis discerned underlying relationships within this group of anticancer drugs and to the parent busulfan. This work shows that pattern recognition methods combined with molecular modeling can discover and elucidate novel drug designs for the clinical treatment of CML.

Keywords: *busulfan, CML, pattern recognition, anticancer, drug design*

Introduction

The term myeloid cell is utilized to designate any leukocyte that is not a lymphocyte and is particularly seen when classifying cancers such as leukemia. Whereas myeloid suggests origin in the spinal cord or bone marrow (or similarity to the bone marrow or spinal cord). Chronic myelogenous leukemia (CML) is a myeloproliferative disorder that is associated with a characteristic chromosomal translocation referred to as Philadelphia chromosome [1]. It is also referred to as chronic myeloid leukemia, chronic myelocytic leukemia, or chronic granulocytic leukemia. Essentially CML is a clonal bone marrow stem cell disease in which there is proliferation of mature granulocytes such as basophils, eosinophils, and neutrophils. The CML condition permits the formation and development of more mature white blood cells and platelets which perform similarly to normal cells in the early stages of the disease. This finding contributes to the often asymptomatic state at actual diagnosis, wherein elevated white blood cell counts are encountered incidental to routine laboratory testing. CML is usually found in middle-aged and elderly individuals, however it can occur in all age groups with a small increase in men. In the U.S. clinical findings show it represents about 20% of all adult leukemia's, and 15% to 20% of all adult leukemia's in Western societies [1]. Notable symptoms include malaise, gout, anemia, low grade fevers, splenomegaly, and thrombocytopenia [1]. Basophils and eosinophils are commonly increased.

Phases of Chronic Myelogenous Leukemia

Utilizing laboratory findings (ie. number of immature leukemia cells (blasts) in blood) and clinical characteristics CML can be divided into phases. The appearance of new chromosomal abnormalities in addition to the Philadelphia chromosome is a decisive feature of progression through the various phases [1]. There is some variation of treatment as based on the concurrent phase.

Chronic Phase

Many patients in this phase can pursue a normal life and are asymptomatic or have modest symptoms of fatigue/abdominal fullness. It is this phase where most individuals are diagnosed incidental to laboratory determination. Approximately 5% (or fewer) of blast cells

are observed in the blood or bone marrow. Treatment can be done as outpatient and includes chemotherapy, chemotherapy with biological therapy (interferon), chemotherapy with donor stem cell transplant, splenectomy, or other drug usage.

Accelerated Phase

In addition to the laboratory finding of 6% to to 30% blast cells, more than 20% basophils, splenomegaly, platelets less than 100000 not related to therapy, the patient will also show infections, fatigue (caused by anemia), and bleeding/bruising. This indicates the disease is progressing. Treatment includes chemotherapy, high-dose chemotherapy, chemotherapy with biological therapy (interferon), stem cell transplant, transfusion, or other drug application.

Blast Crises

In this phase CML appears as an acute leukemia with short survival and accompanied rapid progession [2]. Laboratory findings show development of chloroma, more than 20% myeloblasts/lymphoblasts in blood/bone marrow, and large clusters of blasts in the bone marrow. High mortality is descriptive of this phase. Treatment includes chemotherapy, high-dose chemotherapy, stem cell transplant, chemotherapy to alleviate symptoms, or other drug therapy.

Relapsed Chronic Myelogenous Leukemia

Not universally considered as a phase of CML disease progression, this condition occurs as the number of blast cells increases following a period of disease remission. A molecular remission is determined to be the absence of Philadelphia chromosome. Treatment options include biological therapy (interferon), donor lymphocyte infusion, or donor stem cell transplant.

Busulfan

Busulfan is utilized for the clinical treatment of CML [3]. Busulfan is a cell cycle non-specific less reactive methanesulfonate bifunctional alkylating agent which does not go through a strained ring intermediate but cross-links guanine residues [3], and at the N-7 position [4]. Unlike nitrogen mustards which form interstrand links in DNA, this drug forms intrastrand linkage [4]. Busulfan is eliminated from the body via glutathione conjugation [4]. Busulfan reduces the overall granulocyte mass which reduces the symptoms of CML and improves the clinical situation. Hematologic remission accom-panying stabilization of organomegaly can be attained in as many of 90% of cases. Busulfan treatment is superior to splenic irradiation for favorable results in survival time, maintanence of hemoglobin levels, and found equivalent to controlling spleno-megaly by irradiation. Side effects of busulfan usage includes bone marrow hypoplasia, hyperpigmentation, seizures, and interstitial pulmonary fibrosis.

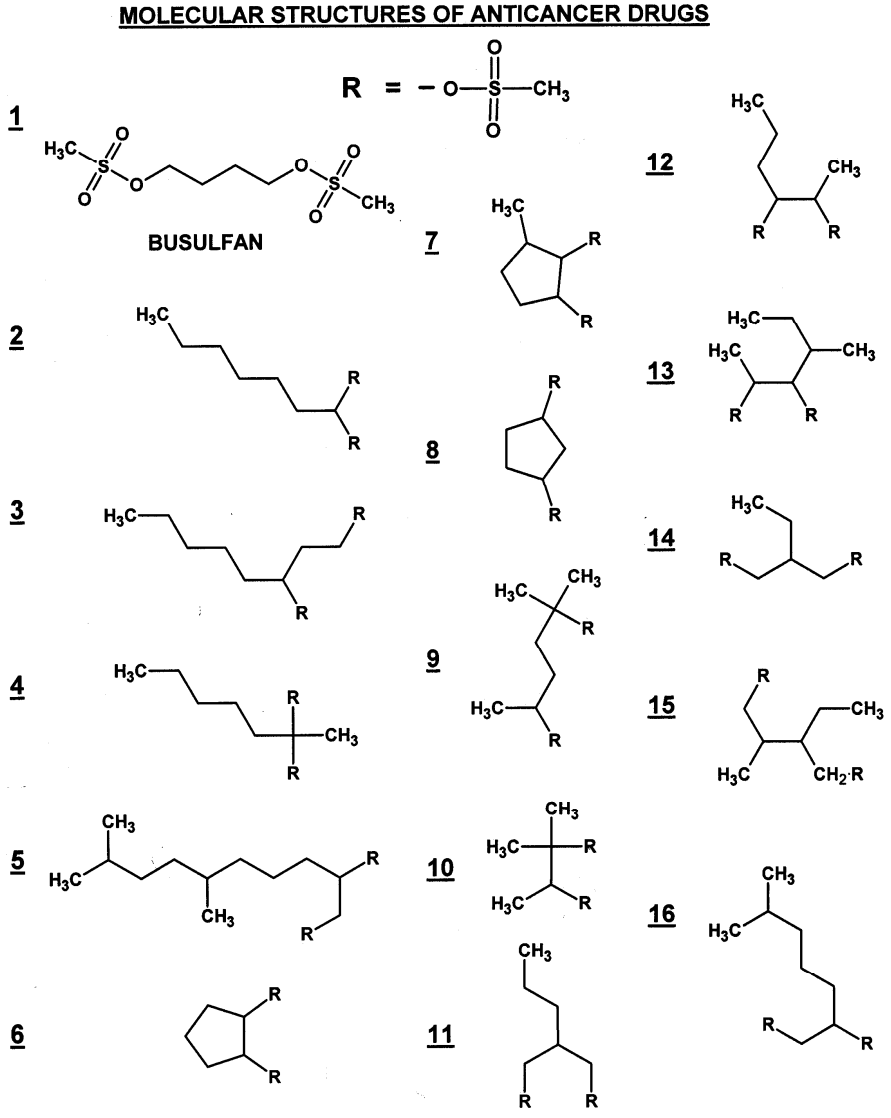

Figure 1. The molecular structure of the parent busulfan (1) is shown for comparison to analog constructs (2 to 16) which differ by modifications including by aliphatic carbon skeleton, branching, and ring moieties. All compounds retain the bifunctional alkylating methanesulfonate substituents.

Rational Drug Design

Rational drug design begins with some information of the pharmacokinetic (absorp-tion, distribution, metabolism, and elimination or ADME) and pharmacodymanic (result of interaction of the drug and body) activity of a drug [5]. Manipulating features of this information can result in an improved fit of the drug for the treatment goals.

In this study the busulfan agent will be considered the parent compound of a family of daughter constructs which retain the methanesulfonate alkylating functional group but show variation in solubility, lipophilicity, polar surface area, molecular volume, molecular weight,

etc due to structural variation of the carbon skeleton. Principals of quantitative structure-activity relationship (QSAR) will be applied to analyze variation of chemical reactivity indicated by changes in the numerical values of important pharmacological properties. Retention of the melthanesulfonate moiety preserves the alkylating and anticancer activity of the daughter constructs.

However the modification of the carbon skeleton will be shown to produce significant changes in lipophilicity, molecular weight, etc, which in turn may benefit the desired clinical goals. Altering the lipophilic nature of a drug in turn can alter and improved bioavailability [5]. In addition, selective modification of the bifunctional structure of busulfan may enhance treatment outcome in the clinical application against CML.

In addition to assimilating numerical values of molecular properties for all drugs in this study, pattern recognition techniques such as cluster analysis, non-metric multidimensional scaling, discriminant analysis, and non-hierarchical K-means cluster analysis will be utilized to discern underlying relationships within this group of anticancer drugs and to the parent busulfan. By accomplishing this it will reveal subtle but very useful pharmaceutical facets enhancing and expanding successful clinical treatment of chronic myelogenous leukemia.

Determination of Pattern Outcome and Molecular Modeling

Molecular Modeling and Determination of Molecular Properties

Determination of molecular properties such as molecular weight and modeling was accomplished utilizing ACD/ChemSketch v. 10.00 (Advanced Chemistry Development 110 Yonge Street, Toronto Ontario, M5C 1T4 Canada). Additional properties; polar surface area, violations of Rule of 5, molecular volume, number of oxygens/nitrogens/amines/hydroxyls, etc were determined by Molinspiration (Molinspiration Chemifor-matics, Nova ulica 61, SK-900 26 Slovensky Grob, Slovak Republic). Determination of isosteric and biosteric substituent analogue elucidation was assisted by utilizing MolSoft molecular analysis software (MolSoft L.L.C., 3366 North Torrey Pines Court, Suite 300, La Jolla CA 92037 USA).

Numerical Analysis and Data Matrix Pattern Elucidation

To determine underlying relationships of tabulated molecular property numerical values various pattern recognition techniques were applied. This includes cluster analysis performed by KyPlot v. 2.0 Beta 15 (copyright Koichi Yoshioka 1997-2001).

Statistical analysis of all numerical data was performed by Microsoft EXCEL (EXCEL 2003, copyright 1985-2003). All remaining pattern recognition determinations (non-metric multidimensional scaling, discriminant analysis, and K-means cluster analysis) were performed by PAST v. 1.80 (copyright Oyvind Hammer, D.A.T. Harper 1999-2008).

Multiple regression was accomplished by GraphPad Instat v. 3.00 for Windows 95 (GraphPad Software, San Diego California USA).

Resolution of Findings and Discussion

Structurally related differences of potency, side effects, and other activities observed among drugs having similarity in structure are considered structure-activity relation-ships (SAR) [5]. Studies of SAR utilizing a parent (lead) compound (busulfan here) and generated constructs (analogs) can help elucidate the component responsible for changes in solubility, polar surface area, etc. The overall goal here is the enhancement of medicinal activity and/or bioavailability and/or ADME characteristics. Previous work has shown that specific approaches to structure variation can produce desirable culmination of druglikeness, which includes [5]: changing size and shape, addition or removal of ring, new substituents, isosteres, and bioisosteres. In outset it is useful to review some aspects of busulfan that support the pursuit of structure alteration. Previous studies have shown that in approximately 90% of CML patients the disease is well controlled by busulfan (or dibromomannitol) [6]. Usually leukemia in association with pregnancy involves CML and findings have shown that busulfan alone is the choice drug for treatment without detrimental effects on the fetus [7]. Hydroxyurea has been compared to busulfan for the treatment of CML in other work [8] and in one study shown to provide extended median survival rates when used alone compared to busulfan [9]. Studies have been done to compare combined application of interferon alpha with busulfan versus hydroxyurea alone [10]. However findings clearly indicate the advantage of busulfan compared to cyclophosphamide in which cyclophosphamide achieved no complete remission of CML and a significant percentage (approximately 30%) of these treated patients endured relapse [11]. The drug Imatinib is a promising targeted therapy of CML which blocks Bcr-Abl tyrosine kinase activity and leads investigators to assert that tailored treatment regimens are possible, however patients can develop resistance to this drug which then requires further intervention by additional chemotherapy [12].

These previous studies and the known toxic side effects of the still effective busulfan clearly purports the pursuit of structural analogs to busulfan to achieve enhanced bioavailability and medicinal action. Utilizing structural renumeration of chemical substituent and analog search (see Molecular Modeling and Determination of Molecular Properties) assisted in focalizing prospective skeletons for developing the family of analogs to busulfan. These analogous compounds are presented in Figure 1 and several general observa-tions of their form can be asserted: 1) Both methanesulfonate functional groups are retained; 2) All variations in structure occur between the two methanesulfonate groups; 3) The modified component are carbon skeletons (ie. no heteroatoms); 4) Three analog compounds have ring structures (6, 7, and 8); 5) All compounds presumably retain bifunctional alkylating capability; 6) Number of oxygens and nitrogens remain the same throughout; 7) Number of non-hydrogen atoms increase over the parent busulfan; and 8) Save for 6, 7, and 8 all analogs have aliphatic or branched aliphatic skeletons. Previous studies have established that introducing different sized rings and chain branching can have effects on potency and activity of an analogous construct [5]. Determination of 11 molecular properties was completed and

are presented in Table 1. A profound change is seen in values of Log P and therefore lipophilicity throughout compared to busulfan. The mean of Log P values is 0.876, median is 0.829, standard deviation is 0.90, range of 3.162, and 95% confidence range is 0.480. The mean of Log P at 0.876 is significantly greater than that for busulfan.

The standard deviation value is 28.5% of the range and is considerably broad. The Log P values of the analogs having rings constituents (6, 7, 8) are -0.087, 0.154, and –0.321, respectively, and have a range of 0.475 which is 15.0% of the population range. This latter finding depicts the variation of lipophilicity property that is possible even though a five carbon ring is common for 6, 7, and 8. Overall the range of Log P is considerable and shows clearly the affect of size and orientation of the carbon skeleton on this important parameter of lipophilicity. Many of the molecular properties remain constant and include the following: polar surface area, number of oxygens and nitrogens, Log BB, number of amine and hydroxyls, BB, and violations of Rule of 5. The mean number of atoms (excluding hydrogens) is integer 17, mode of 17, and a range of 8. The standard deviation value of 2 (integer value) is 25% of the overall range. Likewise the mean for formula weight is 281.872, median is 281.374, range of 112.22, and a standard deviation of 26.379 which is 23.5% of the overall range. Similarly the mean for molecular volume is 234.489, median is 235.20, range of 133.77, and a standard deviation of 32.8 which is 23.9% of the overall range. This findings demonstrate the significant variability of important molecular properties even within analogs to busulfan. All together the apparent variation in these properties can result in considerable alteration in bioavailability which will be discussed below. Contemporary drug development examines large numbers of potential candidates for clinical application within a specific disease aggroup. To expedite and render more effective the selection of candidates having maximal druglikeness one important evaluation process has gained notice and relies on molecular properties to identify. This set of rules known as Rule of 5 pursues properties which most likely enhance oral activity and has considerable importance to considerations of ADME. The Rule of 5 consists of criteria incorporating formula weight, lipophilicity, with hydrogen bond donors and acceptors and describes an orally active drug as follows [13]: 1) There are no more than 5 hydrogen bond donors (nitrogen and oxygen atoms with one or more hydrogen atoms); 2) There are no more than 10 hydrogen bond acceptors (ie. nitrogen and oxygen atoms); 3) A molecular weight of less that 500; and , 4) Partition coefficient Log P less than 5. Looking at Table 1 its seen that all compounds show zero violations of the Rule of 5. This feature is highly advantageous for favorable bioavailability and strongly supports these daughter constructs of busulfan for consideration, remembering all compounds retain the bifunctional methanesulfonate moiety. Hydrophobic interactions are a corresponding interaction of lipohilicity which is conferred often as Log P whereas molecular volume is topological and descriptive of polarizability and van-der-Waals interactions. All compounds presented retain the same number of oxygens and nitrogens so polar surface area remains static. Other studies have shown the considerable importance of polar surface area to successful drug discovery identification. The largest majority of well absorbed drugs are passively transported across lipophilic cell membranes [14]. Structural diversity can seriously impede successful prediction of drug absorption based on partition coefficients [14]. Convincing evidence shows polar surface area (PSA) to be effective in differentiating poorly absorbed drugs when considering oral bioavailability of drug candidates [14]. It is shown that

drugs that are completely absorbed have a PSA of less than 60 Angstroms^2 , whereas drugs having PSA greater than 140 Angstroms^2 are less than 10% absorbed [14]. By these criteria and PSA values it suggests that all analogs of busulfan will have an intestinal absorption of approximately 60%, a high value, that further supports the druglikeness of these analogs.

The prediction of blood-brain barrier (BBB) penetration is an important facet of drug discovery. Whether an investigator desires to penetrate the central nervous system or minimize the brain interaction by a perspective drug candidate it is very useful to assign a numerical value for comparing to agents known to have activity within the central nervous system. The descriptor Log BB (Log [Cbrain/Cblood]) can be estimated by several routes each of which utilized molecular properties considered to have significant influence.

The formula utilized in this work focuses on the PSA contribution and appears as follows [15]: Log BB = -0.016(PSA) + 0.547. Therefore the Log BB value for all compounds is -0.841 (see Table 1) and results in BB values of 0.144. Drugs have a BB value less than 0.1 are poorly distributed in the central nervous system while drugs having BB greater than 2.0 can readily cross the BBB. The analogs to busulfan presented here would not be expected to have substantial partitioning across the BBB. Other studies have shown that small molecules have improved tendency to penetrate the BBB [16], while other work has focused on the combined influence of drug lipophilicity as represented by Log P descriptor. This latter study claims optimal requirements for crossing the BBB are fulfilled via a Log P value between 1 and 4, a molecular weight less than 400, with a PSA value less than 90 Angstroms^2 [17].

Table 1. Molecular Properties of Anticancer Drugs

DRUG	LOG P	POLAR SURFACE AREA	NUMBER OF ATOMS	FORMULA WEIGHT	NUMBER OF O & N	NUMBER OF ROTATABLE BONDS	MOLECULAR VOLUME	LOG BB	BB	NUMBER OF -NH AND -OH	VIOLATIONS OF RULE OF 5
BUSULFAN 1	-0.428	86.752	14	246.306	6	7	193.796	-0.841	0.144	0	0
2	1.959	86.752	17	288.387	6	9	244.987	-0.841	0.144	0	0
3	1.737	86.752	18	302.414	6	10	260.789	-0.841	0.144	0	0
4	1.900	86.752	17	288.387	6	8	243.422	-0.841	0.144	0	0
5	2.734	86.752	22	358.522	6	12	327.566	-0.841	0.144	0	0
6	-0.087	86.752	15	258.317	6	4	199.808	-0.841	0.144	0	0
7	0.154	86.752	16	272.344	6	4	216.395	-0.841	0.144	0	0
8	-0.321	86.752	15	258.317	6	4	199.808	-0.841	0.144	0	0
9	0.745	86.752	17	288.387	6	7	243.207	-0.841	0.144	0	0
10	0.204	86.752	15	260.333	6	5	209.603	-0.841	0.144	0	0
11	0.839	86.752	16	274.360	6	8	227.185	-0.841	0.144	0	0
12	0.819	86.752	16	274.360	6	7	226.970	-0.841	0.144	0	0
13	1.006	86.752	17	288.387	6	7	243.557	-0.841	0.144	0	0
14	0.280	86.752	15	260.333	6	7	210.383	-0.841	0.144	0	0
15	1.026	86.752	17	288.387	6	8	243.772	-0.841	0.144	0	0
16	1.447	86.752	18	302.414	6	9	260.574	-0.841	0.144	0	0

Polar Surface Area = Angstroms^2
Molecular Volume = Angstroms^3
Log BB = Log (Cbrain/Cblood)
BB = Cbrain/Cblood
Note that number of atoms count ignores hydrogens.

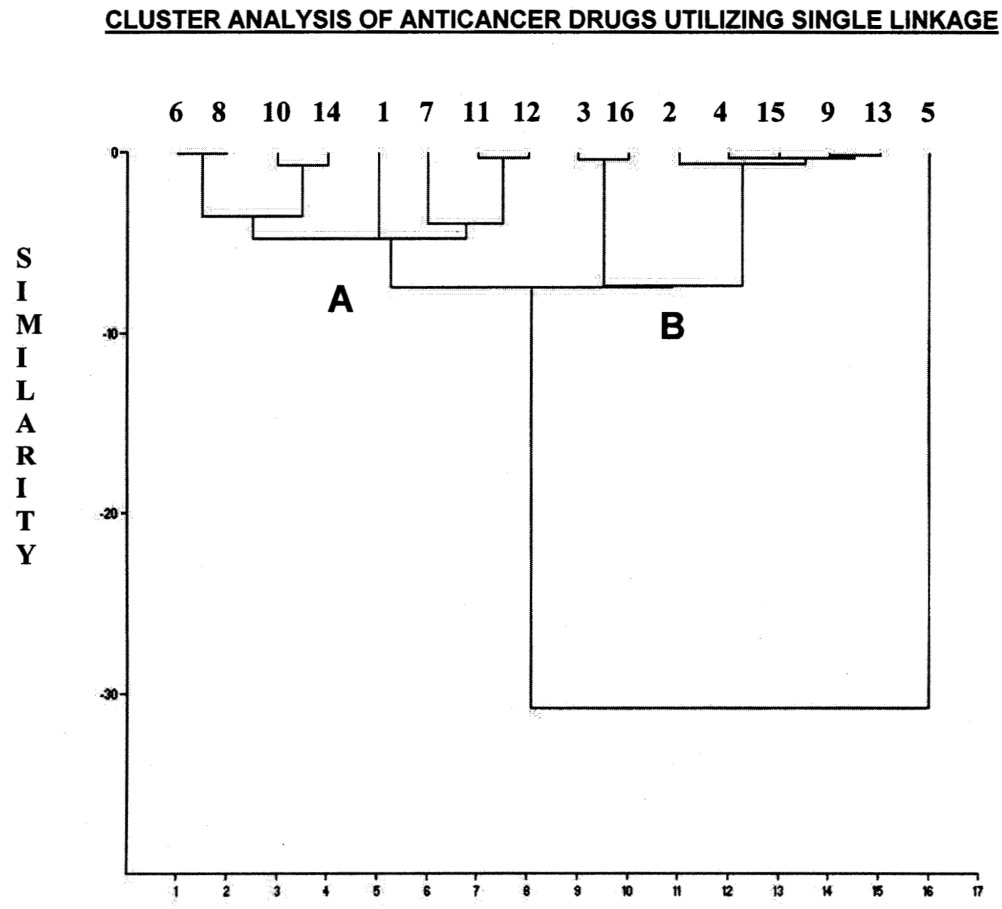

Figure 2. Hierarchical cluster analysis is performed on drug constructs presented in Figure 1 utilizing standard Euclidean distance and single linkage conditions. Node A join constructs 6 and 8, 10 and 14, 1, 7, 11 and 12. Node B joins 3 and 16, 2, 4 and 15, 9 and 13. Analog 5 is viewed as distinct from busulfan (1) and the remaining analog population.

This latter criteria is fulfilled by analogs 2, 3, 4, 5, 13, 15, and 16 (see Table 1), which suggests these analogs of busulfan may be useful in the treatment of malignant tumors of the brain. Currently it is estimated that 13000 deaths per year result from brain tumors [18] and in 2005-2006 they accounted for 20% to 25% of pediatric cancers [19]. There-fore this potentially beneficial spin-off shows the efficacy of deriving analogs from an established clinical medication. The Pearson r product-moment correlation relates the strength of linear relationship, magnitude, and direction of association among variables. Pearson correlation analysis was performed on the data matrix of Table 1, producing Pearson r greater than 0.9900 for all 16 drugs. That is, by these molecular properties all 16 drugs showed maximal positive correlation with all other members. Analysis of similarities was performed upon the data matrix of Table 1. Analysis of similarities (ANOSIM) is a statistical analysis to determine whether a significant, difference exists between two or more groups of data [20]. A positive value up to value of one indicates significant difference between groups. Results of analysis by ANOSIM on Table 1 produced a value of 0.160, which indicates a significant level of similarity among these analogs and parent busulfan.

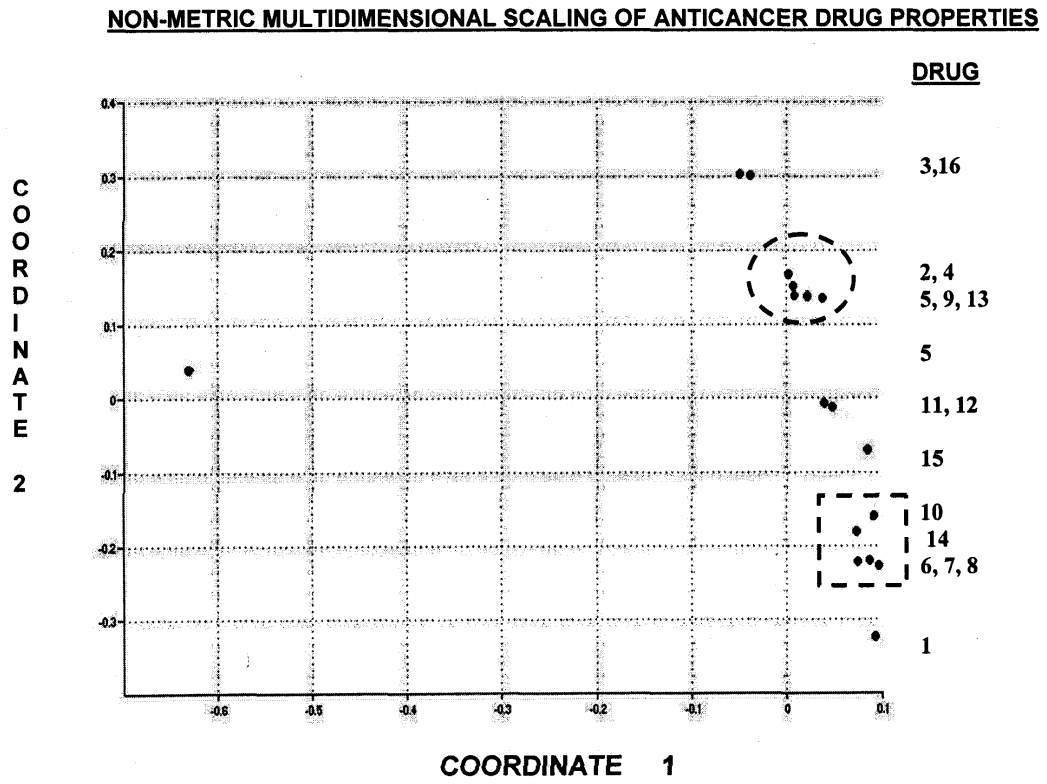

Figure 3. Non-metric multidimensional scaling preserves the ranked differences within the data matrix while presenting the results in a 2-way coordinate system. Drug 5 is determined to be substantially distinct from all remaining constructs. Analog constructs 2, 4, 5, 9, and 13 (enclosed circle) are grouped altogether as are analogs 10, 14, 6, 7, and 8 (enclosed square). Note that the parent busulfan (1) is determined to be distinct from both of these largest groupings.

The purpose of discriminant analysis (DA) is to determine the combinations of independent variables which best discriminates among the given groups (molecular properties of the drugs presented here) [21, 22]. Observations include arrangement of the sets to maximize differences [22]. Cluster analysis is not identical to DA because clusters (groups) are determined in advance prior to assignment of independent variables that discriminate groups. Applying DA to the data matrix of Table 1 produced two groups having maximal differences based on molecular properties: Group A) Bulsulfan (drug 1), 3, 6, 8, 11, 14; and Group B) Drugs 2, 4, 5, 7, 9. 10, 12, 13, 15, and 16. Interestingly all analogs, save for number 3, which are suitable to cross the brain-blood barrier (analogs 2, 3, 4, 5, 13, 15, and 16) fall within Group B. Busulfan, the parent compound, falls within Group A and is determined to be distinct from BBB crossing Group B members as well. This finding clearly shows the enhancement of drug bioavailability possible by structure modification over even the parent compound. In addition, employing the DA algorithm will differentiate members of a population of compounds, via molecular properties, into groups that suggest unvariedness among peers. Assembling subjects (drugs) within a data matrix according to similarity of members within the determined clusters is accomplished by K-means cluster analysis. This method differs from hierarchical clustering in some important aspects [23]: 1) There is

hierarchy and the data is partitioned; 2) There is no dendrogram and only the final cluster membership for each case is presented; 3) The investigator supplies the numbers of clusters (k) into which the data are to be grouped. In a simple presentation of steps the process occurs as follows: Step 1) All cases are initially assigned randomly into the designated number of clusters; Step 2) Cases are displaced among clusters in an iterative manner so that clusters are internally similar but externally dissimilar to other clusters; Step 3) Movement is discontinued when displacement between clusters makes the clusters become more variable. To ascertain levels of similarity among the 16 compounds of Table 1, begin with assigning 3 cluster for K-means results and obtain the following: Cluster A) 1, 6, and 8; Cluster B) 2, 3, 4, 5, 9, 13, 15, and 16; and Cluster C) 7, 10, 11, 12, and 14. Interestingly all analogs determined to pierce the BBB by Log P, molecular weight, and PSA (analogs 2, 3, 4, 5, 13, 15, and 16) are found in Cluster B and have higher similarity within a total number of three clusters. Busulfan is found to have greater similarity to analogs 6, and 8. Increasing resolution into a total of four clusters produced the following results: Cluster A) 1 alone (busulfan); Cluster B) 6, 7, 8, 10, and 14; Cluster C) 2, 3, 4, 9, 11, 12, 13, 15, and 16; and Cluster D) 5 alone. Save for analog 5 all BBB crossing analogs persist within the same group Cluster C. Busulfan is now visualized as distinct from all analog drugs 2 thru 16. Still higher resolution of this population of alkylating anticancer drugs into 5 clusters yields the following results: Cluster A) 11 and 12; Cluster B) 1, 6, 8, 10, and 14; Cluster C) 5 only; Cluster D) 7only; and Cluster E) 2, 3, 4, 9, 13, 15, and 16. At this level then busulfan is shown to be closer to 6, 8, 10, and 14. All BBB crossing drugs, save for number 5, are still most similar and fall into the same Cluster E. The distinction of busulfan into an unshared private cluster by four groupings suggests that level of analysis is optimal (ie. In that the parent compound is shown to be individual and separated from all analogs of the population). Multiple regression establishes a mathematical relationship among independent variables and a dependent variable, and this study various descriptors will be utilized to define a useable equation for calculating the formula weight of analogous compounds. Namely Log P, number of atoms (nAtoms) ignoring hydrogens, number of rotatable bonds (nRotB), and molecular volume (MV) will be treated as the independent variables and formula weight (FW) the dependent variable. The equation formed is as follows: FW = 58.067 + 0.06222(Log P) + 10.511(nAtoms) + 0.07441(nRotB) + 0.2146(MV). This equation provides an R squared of 100.00% which is the level of variance explained by the model (P value less than 0.0001). Most significant influence is contributed by number of atoms, number of rotatable bonds, molecular volume, and constant 58.067. This equation can be utilized to postulate bifunctional methanesulfonate constructs resembling this population. Hierarchical cluster analysis produces a dendrogram showing how data subjects (drugs in this study) can be clustered in which subjects within any cluster are more similar to each other than to subjects found in separate clusters [21]. A vertical dendrogram (hierarchical) is presented in Figure 2 showing results of cluster analysis of molecular properties from Table 1. Hierarchical analysis allows multiple joining of subjects to individual nodes. Strikingly analog 5 is shown to be highly distinct from all remaining drugs including the parent busulfan. Drug 5 possesses the longest aliphatic carbon chain (branched) than any other analog of the population and may be responsible for its distinctness identified by hierarchical cluster analysis and four cluster non-hierarchical K-means cluster analysis. Two super clusters are formed having in turn analogs 3, 16, 2, 4, 15, 9, 13, 15 (joined at node

B, see Figure 2) and 6, 8, 10, 14, 1 (busulfan), 7, 11, and 12 (joined at node A). Each super cluster is divided into subclusters which identify individual or multiple subjects as having greatest similarity. The analysis utilized simple Euclidean distance (the shortest distance between subjects in the multivariate space) and single linkage between clusters which is the distance between the two most closest subject contained within the clusters. Interestingly busulfan (drug 1) is shown to be distinct from other subjects joined at node A. Several subjects are paired such as 6 and 8, 10 and 14, 11 and 12 but unique from 1 and 7 (not paired), which is a result that resolves the parent busulfan from the analogs (remembering that busulfan clearly falls in common at node A). Super cluster at node B joins 3, 16, 2, 4, 15, 9, and 13 which are also broken into pairs as subclusters save for 2 which is considered distinct within this group of analogs. Paired subjects joined at node B are 3 and 16, 4 and 15, 9 and 13. Some level of similar medicinal activity could occur for the indicated paired drugs based on molecular properties. Multidimensional scaling (MDS) is an analysis with the goal of detecting meaningful underlying dimensions that explain observed similarities and dissimilarities [21]. This type of pattern recognition has great flexibility in accepted data and can analyze any kind of similarity or dissimilarity matrix as well as correlation matrix.

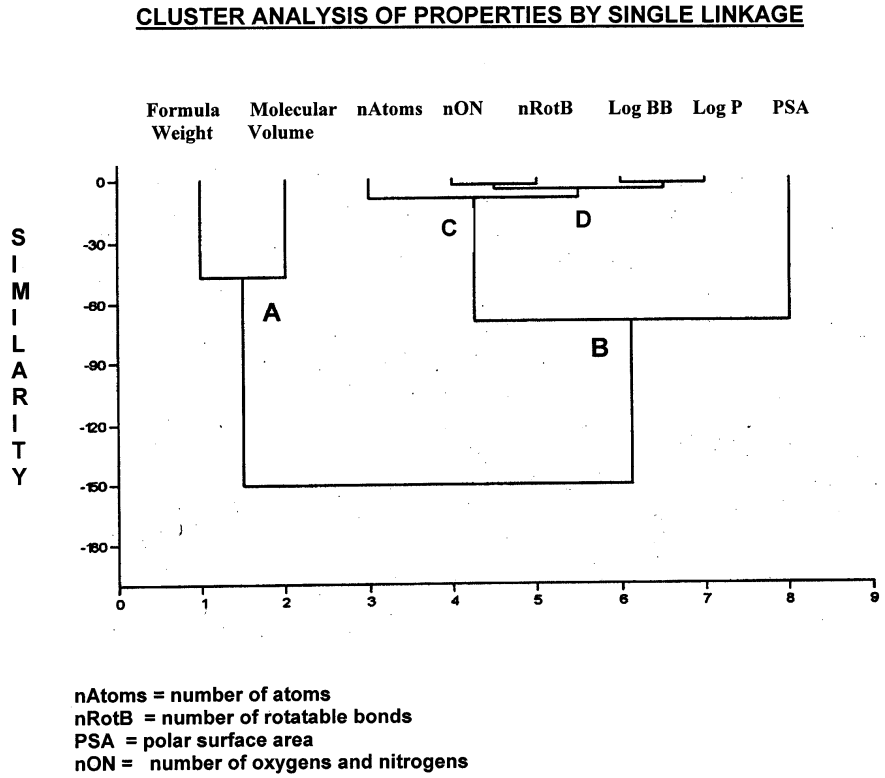

Figure 4. Hierarchical cluster analysis of descriptors reveals associations among the properties not appearing through visual inspection of numerical values and K-means cluster analysis. Clearly at node A formula weight and molecular volume are paired and distinct from all other descriptors. Node B coalesces all remaining descriptors however further resolution is obtained by differentiating polar surface area and joining the remainder (save for number of atoms) at node D.

The analysis attempts to arrange subjects into dimensions that reproduce their distances; in which two dimensions have been found to be highly useful and flexible. There are two types of MDS, metric and non-metric, wherein the non-metric model has fewer restrictions and less rigor. Non-metric multidimensional scaling is performed on the data matrix of Table 1 with results shown Figure 3. The algorithm places the subjects (drugs) into the two-dimensional coordinate system in a manner that preserves their ranked differences. Therein the drugs are arranged so that closest neighbors are considered more similar (based on the tabulated molecular properties) and distance measurements accomplished in the manner like cluster analysis. Presentation of results are seen in Figure 3 along axis coordinate 1 and coordinate 2. Strikingly it is seen that analog 5 is a relatively large distance from all the remaining drugs including the parent busulfan, an outcome due to its long and branched aliphatic carbon chain. However, the parent busulfan (drug 1) is also distinguishable from all other drugs as it lies a significant distance along coordinate 2 from a grouping that in includes 10, 14, 6, 7, and 8 (also joined together by cluster analysis at node A). Another grouping occurs higher along coordinate 2 and includes 2, 4, 5, 9, and 13 (save for 5, these also shown similar by cluster analysis at node B). Satellite paired subjects are 3 and 16 (these also paired by cluster analysis; see Figure 2), as well as 11 and 12 (these also paired by cluster analysis). As before these results suggests that analogs seen as similar in the two-way coordinate system may demonstrate alike bioavailability and ADME characteristics.

Elucidation of Property Relationships

Correlation and underlying interactions of the following descriptors can be achieved also by pattern recognition scrutiny: Log P, polar surface area, number of atoms (excluding hydrogens), formula weight, number of oxygens-nitrogens-roratable bonds, molecular volume, and Log BB. While the numerical values of these descriptors vary widely the pattern recognition techniques determine interrelationships that are not obvious to superficial inspection. Beginning with non-hierarchical K-means cluster analysis and determination into merely two clusters will separate the descriptors as follows: Cluster 1) Log P, polar surface area, number of atoms, nitrogens, oxygens, roratable bonds, Log BB; and Cluster 2) Formula weight and molecular volume. Although the numerical values of formula weight and molecular volume appear logical, note that for the descriptor of Cluster 1 the percent difference in the raw numbers is far greater, yet K-means consider these properties similar to the extent of identifying two groupings. Repeating the K-means analysis into four initial clusters produces higher resolution of the descriptors and the appearance of substantial influence due to the raw numerical values. The results show four clusters as follows: Cluster 1) Polar surface area; Cluster 2) Number of atoms, number of oxygens and nitrogens, number of rotatable bonds; Cluster 3) Log P, Log BB; and Cluster 4) Formula weight and molecular volume. This non-hierarchical method will not show inter-cluster associations (if any), however hierarchical cluster analysis will provide some insight into this possibility. Hierarchical cluster analysis results for descriptors of Table 1 are shown in Figure 4. Hierarchical analysis will show inter-cluster association in addition to grouping into clusters based upon similarity within the data matrix. On the outset it is readily seen that formula

weight and molecular volume are paired into a cluster that is distinct from all other descriptors and joined at node A. All other descriptors are joined at node B and further differentiated into clusters. Of these descriptors the algorithm views polar surface area as distinct from the remaining and placed into a unique cluster. Node C joins the individual number of atoms to node D which has two subclusters of paired number of rotatable bonds and number of O and N, with paired Log BB and Log P. Overall this hierarchical analysis views all descriptors save for formula weight and molecular volume to have some level of inter-association. Con-sidering node B only then, further resolution of the remaining descriptors is accom-plished by separating polar surface area from the balance and node C is divided into node D and number of atoms. The analysis views Log BB, Log P, nRotB, and nON to have association but separable into pairs. These observations contribute to the understanding of the association of properties to realization of bioavailability and ADME characteristics.

Conclusion

Chronic myelogenous leukemia is a myeloproliferative disorder that represents about 20% of all adult leukemia's, and 15% to 20% of all adult leukemia's in Western societies Busulfan is bifunctional alkylating methanesulfonate drug that reduces the overall granulocyte mass and reduces the symptoms of CML. Side effects of busulfan includes bone marrow hypoplasia and seizures. Rational structural modification can vary the property characteristics and enhance bioavailability and ADME attributes. Notable outcomes of property determination reveal that all analogs and parent busulfan show zero violations of the Rule of 5, suggesting favorable bioavailability. Indicator of potential central nervous system interaction are suggested be values of Log BB which remains constant at -0.841 (BB values constant at 0.144). Utilizing the criteria of Log P value between 1 and 4, a molecular weight less than 400, and PSA value less than 90 Angstroms^2 suggests that analogs 2, 3, 4, 5, 13, 15, and 16 can cross the blood-brain barrier and presumably assail tumors of the central nervous system. A broad range in Log P values are obtained (from -0.321 to 2.734) due to the aliphatic, branched aliphatic, and ring substituents of the constructs. The number of oxygens, nitrogens, and polar surface area remain constant throughout the entire population of constructs at 6, 0, and 86.752 A^2, respectively. Pattern recognition techniques provide information concerning underlying associations that exist among subjects of a numerical data matrix. Hierarchical cluster analysis revealed analog 5 to be highly distinct from all other analogs and parent busulfan due to the long branched aliphatic carbon skeleton positioned between two methane-sufonate substituents. Further elucidation of drugs 1 to 16 discerned two superclusters having analogs, respectively: Node A) 6, 8, 10, 14, 1 (busulfan), 7, 11, 12 ; and Node B) 3, 16, 2, 4, 15, 9, 13, and 5. Non-hierarchical K-means cluster analysis also orientates subjects showing similarities but no hierarchical associations. Non-metric multidimensional scaling analyzes multivariate data matrix and displays outcomes in a 2-way coordinate system that preserves ranked differences. Results here show analog 5 to be highly distinct and with two large clusters forming: (A) 2, 4, 5, 9, 13; and (B) 10, 14, 6, 7, and 8. Hierarchical cluster analysis of properties showed associations not immediately apparent from

raw number examination. Namely formula weight and molecular volume are most similar and highly distinguishable from Log BB, Log P, nRotB, and nON, all of which are joined into a superior node distinct from polar surface area and nAtoms. This study showed clearly the substantial efficacy of rational drug design utilizing pattern recognition techniques and modeling structural analogs from a generative pharmaceutical anticancer agent, for this work, busulfan.

Acknowledgements

This work was supported by the University of Nebraska, College of Arts and Sciences, Department of Chemistry, 6001 Dodge Street, Omaha NE 68182, USA.

References

[1] Faderl, S; Talpaz, M; Estrov, Z; Kantarjian, HM. (1999). Chronic mylogenous leukemia: biology and therapy. *Annals of Internal Medicine*, 131(3), 207-219.

[2] Tefferi, A. (2006). Classification, diagnosis and management of myeloproliferative disorders in the JAK2V617F era. *Hematology Am. Soc. Hematol. Educ. Program*, 240-245.

[3] Gringauz, A. Medicinal Chemstry. New York: Wiley-VCH; 1997.

[4] Silverman, RB. The Organic Chemistry of Drug Design and Drug Action. San Diego:Academic Press; 1992.

[5] Thomas, G. Medicinal Chemistry. New York: Wiley; 2000.

[6] Huguley, C. (2006). Chronic myelocytic and chronic lymphocytic leukemia. *Cancer*,30(6), 1583-1587.

[7] Nolan, GH; Facog, RM; Perez, C. (1971). Busulfan treatment of leukemia during pregnancy. *Obstetrics and Gynecology*, 38, 136-138.

[8] Rushing, D; Goldman, A; Gibbs, G; et al. (1982). Hydrea vs busulfan in the treatment of chronic myelogenous leukemia. *Am. J. Clin. Oncol*, 5, 307-305.

[9] Bolin, RW; Robinson, WA; Sutherland, J: Hamman, RF. (2006). Busulfan versus hydroxyurea in long-term therapy of chronic myelogenous leukemia. *Cancer,* 50(9), 1683-1686.

[10] Hehlmann, R; Heimpel, H; Huford, J; et al. (1994). Randomised comparison of interferon alpha with busulfan and hydroxyurea in chronic myelogenous leukemia. *Blood,* 84, 1064-1067.

[11] Kaung, DT; Close, HP; Whittington, RM; Patno, ME. (2006). Comparison of busulfan and cyclophosphamide in the treatment of chronic myelocytic leukemia. *Cancer*, 27(30), 608-612.

[12] Jabbour, E; Cortes, JE; Ghanem, H; O'Brien, S; Kantarjian, HM. (2008). Targeted therapy in chronic myeloid leukemia. *Expert Rev. Anticancer Ther*, 8(1), 99-110.

[13] Lipinski, CA; Lombardo, F; Dominy, BW; Feenye, PJ. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Del. Rev*, 46, 3-26.

[14] Palm, K; Stenberg, P; Luthman, K; Artursson, P. (1997). Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research,* 14, 568-572.

[15] Clark, DE. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration. *Journal of Pharmaceutical Sciences*, 88(8), 815-821.

[16] van de Waterbeemd, H; Kansy, M. (1992). Hydrogen-bonding capacity and brain penetration. *Chimia,* 46, 299-303.

[17] van de Waterbeemd, H; Camenisch, G; Folkers, G; Chretien, JR; Raevsky, OR. (1998). Estimation of blood-brain crossing of drugs using molecular size and shape, and H-bonding descriptors. *Journal of Drug Targeting*, 6, 151-165.

[18] Greenlee, RT; Murray, T; Bolden, S; Wingo, PA. (2000). Cancer statistics. *CA Cancer J. Clin,* 50, 7-33.

[19] Chamberlain, MC; Kormanik, PA. (1998). Practical guidelines for the treatment of malignant gliomas. *West J. Med*, 168, 114-120.

[20] Clarke, KR. (1993). Non-parametric multivariate analysis of changes in community structure. *Australian Journal of Ecology*, 18, 117-143.

[21] Duda, RO; Hart, PE; Stork, DG. Pattern Classification. 2nd Edition. New York: Wiley; 2001.

[22] Davis, JC. Statistics and Data Analysis in Geology. New York: John Wiley and Sons; 1983.

[23] Bow, ST. Pattern Recognition. New York: Marcel Dekker; 1984.