A Computer Vision-Based Approach For Answering Molecular Similarity Queries

Rahul Singh
School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, Georgia 30332

ABSTRACT

This paper presents a method for answering the problem of molecular similarity. This problem is fundamental to the process of drug discovery as well as in structural biology. In the proposed approach the similarity problem is considered in a setting where the molecules are defined using complex surface-based representation. Such representations have a very high descriptive power but have rarely been used in similarity queries due to computational complexity. Given the surface based description of a molecule, the essence of our approach consists in encapsulating it inside a tessellated sphere, at the surface of which we conduct measurements of the molecular fields and geometry. These measurements constitute distributions that describe the molecule. The answer to similarity queries is then obtained by computing a histogram intersection of different distribution for the query and the model molecules. The need for explicit pose optimization in 3D is ameliorated by the use of a novel topological constraint to weight the intersection scores. Among other advantages, the method can deal with super-positioning field effects and facilitates rapid determination of similarity. Its efficacy is demonstrated in terms of its recognition performance and through comparisons with existing research and commercial approaches.

1. INTRODUCTION

The sequencing of the human genome has been a significant achievement in science [7]. It promises fundamental progress not only towards understanding the process of life but also towards the development of novel therapeutics, and eradication of diseases. This has also spurred research on various problems in life sciences both from within that domain and also from other disciplines of science and engineering. The work presented in this paper constitutes an example of the latter type of research. We present a method motivated by research in computer vision, to address the problem of determining molecular similarity, which is fundamental to many biological investigations as well as in drug discovery. In addition to its links to Physics and Chemistry, this problem is also inextricably linked to the fundamental questions of pattern analysis and recognition (see [6], [14], [17] for an introduction) and provides numerous opportunities for inter-disciplinary research as well as for deepening our understanding in these areas. Determining molecular similarity is important because in general similar molecules tend to behave similarly [5]. Evidence of how this observation dictates scientific investigations abound: Biologists use classes of structurally similar molecules to probe proteins or enzymes for binding. In drug discovery, for instance, similar molecules tend to show similar behavior in terms of effects like absorption, distribution, or toxicity. The similarity of a molecule with a known class of molecules can therefore be an important indicator of its potential suitability in an investigation. Furthermore, molecules that are dissimilar to known classes of molecules indicate novel regions of the chemical space and can be candidates for new or more effective therapeutics. For a computational solution, the problem of molecular similarity presents some unique issues that need to be considered:

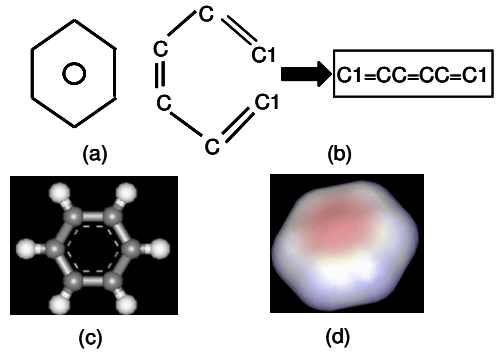

Figure 1: Representations for the Benzene molecule: (a) graphical, (b) string, (c) 3D (ball and stick), (d) surface-based in 3D

- *Representation, Pose, Deformations*: As complex 3D shapes, molecules need to be represented adequately and in a manner that is consistent for explaining their biological behavior. During a similarity query, the pose of a participating molecule can be arbitrary. A molecule can also deform, i.e. take one of many energetically minimal configurations (called conformations). Accounting for this is important because the bio-chemical behavior of a molecule can vary significantly depending on its conformation.
- *Multi-modal nature of representation*: Properties like geometry and charge distributions that may be used for describing molecules have different characteristics. For

example, while geometric representations should be unique, field-effects are superposition-based. Thus, structurally different molecules may show similar biological activity (due to similar field-effects). A similarity method should account for this.

- *Query efficiency*: It is typical to conduct molecular similarity queries over large sets (thousands to millions of molecules). Similarity approaches should therefore be computationally efficient.

The proposed method approaches the problem in three steps: First, the molecule is encapsulated inside a tessellated sphere and its geometry and field properties are computed at the tessellate points in a manner that accounts for their distinct characteristics. The molecule can then be treated as a collection of distributions defined on the sphere, where each distribution represents a specific property. The second step is based on the idea that the similarity of two molecules can be obtained by comparing the similarity of the respective distributions. For two distinct molecules, each in a specific conformation, there may be multiple points on the respective spheres that have the same value for a given property. However, the relative geometric distribution of these points will differ for every molecule. One way to estimate this difference is by considering the distribution of pair-wise distances between these points. This distribution is invariant to the pose of the molecule. The similarity of these distributions can be used as a constraint to determine the similarity of the property distribution. Essentially, high scores are obtained by similar property distributions, only if the distance distributions also tend to agree. An advantage of this approach is that we can formulate its implementation using a topologically-constrained variant of histogram intersection. The technique of histogram intersection [20] has been extensively used in image-retrieval for rapid querying over large data sets. We seek to take advantage of this efficiency. To make the constraint we described above invariant to conformations, in the third and final step we represent a molecule by generating a set of energetically minimized structures, over which the similarity is computed.

While this work specifically focuses on *small molecules* (tens of atoms, molecular weight around one thousand Daltons), typical of the type used in drug discovery, the technique presented here can scale-up equally well to larger molecules like proteins.

We begin this paper with an introduction to molecular representations (Section 2). In Section 3 we present an overview of the prior research in the area. The proposed method is formulated in Section 4. In this section we also relate this approach with work done in Computer Vision using spherical representations (EGI, complex-EGI, and SAI). Experimental results are reported in Section 5. The conclusions and directions of possible future work are presented in Section 6.

2. MOLECULAR REPRESENTATION

As is well known in pattern recognition, the concept of *similarity* is tied to the issues of *representation* and *similarity measure*. This is also reflected in molecular modeling research where formalisms having different representation power have been developed. In the simplest form, a molecule may be represented by using its chemical formula. Some other ways of describing molecules are illustrated in Figure 1, using the Benzene molecule as an example: (a) is the graphical representation of Benzene. The molecular graph is a connectivity matrix where atoms that participate in the bond are shown to be connected. This graph can also contain information about bond orders and can be used to distinguish isomers (same molecular formula but different topologies). In (b) a string representation called SMILES is shown. It is obtained from a depth-first traversal of the molecular-connectivity-graph and represents a molecular valance model. An important advantage is its efficiency for storage and processing. (c) presents a 3D graph (ball-and stick model), that represents an energetically stable configuration of the molecule in terms of the atom positions, the inter-atomic distances, the bond lengths, and the angles between various bonds. Such a model, in addition to having the descriptive power of the previous schemes, can also describe molecular conformations. Finally, (d) shows a surface-representation of a molecule obtained by rolling a probe-atom over the molecule. The molecular surface being defined as the set of points where the surface of the probe atom touches the *van der Walls* surfaces of the atoms constituting the molecule. The surface-based representation can better reflect phenomena that happen across the molecular interface, like molecule-molecule interactions. Moreover, field effects like charge distributions, polar surface area, or donor (acceptor) fields can be defined over the molecular surface, thus providing a better representation of the actual physics as opposed to idealized representations like the ball-and stick model. It should also be noted that the most accurate way to model a molecular system is through the use of the Schrodinger wave equation (1):

$$E\psi = H\psi, \psi = \psi(t) \tag{1}$$

$$i\frac{h}{2\pi}\frac{d}{dt}\psi = \frac{-\left(\frac{h}{2\pi}\right)^2}{2m}\nabla^2\psi + V\psi \tag{2}$$

In this equation, E denotes the energy of the system, H is a self-adjoint operator in Hilbert space, called a Hamiltonian and Ψ encodes the probability of the outcome of all possible measurements made on the system. However, as equation (2) shows, even for a trivial system

consisting of a single particle of mass *m,* the Schrödinger equation is a non-linear PDE. Due to the computational complexity associated with solving such systems, the use of the Schrödinger equation is precluded in most real-world problems where it is common to work with thousands of molecules, each having multiple atoms.

3. MOLECULAR SIMILARITY AND PRIOR RESEARCH

A variety of formulations for molecular similarity can be found in the literature an overview of which may be obtained from [14]. With respect to the context of this paper, we consider three classes of approaches: those that use a 2D molecular graph, techniques based on the 3D molecular graph, and techniques that employ surface-representations. Of these due to their simplicity, 2D graph-based techniques have been very popular. Most successful amongst these are techniques that compute a linear descriptor from the 2D graph, encoding the connectivity and physico-chemical attributes as a bit string. For similarity the bit strings are compared using Euclidean, Hamming, or Tanimoto distance. Such approaches [23] have been used to build commercial systems that can query millions of molecules rapidly. While 2D similarity is adequate for many problems, it typically does not suffice in situations where effects involving molecular shape or field-effects are involved. 3D molecular searching has therefore been an active area of work. Early works in the area like [19] and [22] used variations on the sum of inter-atomic distances. Later approaches have looked at schemes for atom re-labeling to minimize a difference-distance matrix [1], [16] or decomposing the molecular distance and connectivity graphs into subgraphs which are numerically characterized and compared [2]. Another class of methods [8], [11], [12] has defined molecular similarity using surface and field characteristics. First, the field-effects around a molecule are estimated. Then the orientation of the query or model molecule (a 3D graph), is varied to minimize an RMS error between the field values. Other efforts include the application of *geometric hashing* and its variations [15], [18]. Our approach differs from these works in many respects: Unlike pure geometric techniques, we consider geometry as well as field-based descriptors. We also do not suffer from the problem of isospectrality in graphs that is inherent in many molecular graph-based methods [13]. Methods that use geometric hashing are typically formulated around sub-structures of a molecule in 3D, while we consider the problem of whole-molecule similarity. Unlike [8], [11], [12], our approach does not require us to systematically vary the pose of the model (query). Not only does this make the proposed technique much faster, but we also do not suffer from potential errors due to alignment or effects of step-size during variation of orientation.

4. THE PROPOSED METHOD

4.1. The Molecular Representation

The molecular description is computed by placing the molecule inside a semi-regularly tessellated sphere *S,* such that its center of mass coincides with the center of the sphere (Figure 2(a)). Our approach for tessellating the sphere is based on [9] and involves subdivisions of the triangular sides of a 20 side icosahedron into sub-triangles. Let n denote the subdivision frequency and T the number of triangles, let also R denote the radius of the sphere and G denote the molecular graph. Then, the tessellation process can be described as follows:

$$T = 20n^2, n = 10 \quad (3)$$

$$R = radius(G) + 2 \quad (4)$$

At each point of the tessellated sphere we compute three properties to describe the molecule, namely *geometric shape*, *donor field* (due to H-bond donor atoms), and *acceptor field* (due to H-bond acceptor atoms). Our selection of these descriptors is due to their importance in various molecular interactions and the fact that more complex descriptors like polar surface area of the molecule (which influences membrane permeability), are correlated to donor/acceptor fields [21]. The issues encountered in representing molecular shape are similar to those faced in curved object recognition using techniques based on the Gauss map (EGI and complex-EGI). The main problem with the EGI family of techniques is that two or more parts of an object surface may get mapped to the same point on the sphere. The SAI [9] addresses these problems. However, its use in our situation is precluded by the speed issues associated with the convergence of the deformable surface used for mapping an object to a sphere. We present a novel alternate by first noting that the molecular surface, by its construction is *smooth*. At each point P_j of *S,* this surface is probed to determine the surface point closest to P_j that has not already been measured by any point P_k adjacent to P_j. This acts as a local regularity constraint. The distance to the surface point is then used as an estimate of molecular shape. Figure 2(b) illustrates this process.

The measurement of the donor field is done using the following three step procedure:

- The H-bond donor atoms in the molecule are identified. Typically these are Nitrogen or Oxygen atoms with hydrogen on them. Other ways of identification like the PATTY-rule [3] can also be used.

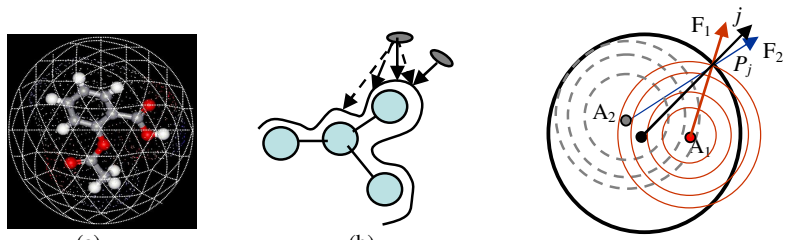

Figure 2: (a) The embedded molecular surface, (b) measurement of the molecular shape, (c) field measurement at a point

- The donor field is defined as an isotropic Gaussian distribution and the field at point P_j due to an atom at position X_i having van der Walls radii r_i is defined as:

$$f(P_j, X_i) = \left(\frac{a^2}{2\pi r_i^2}\right)^{\frac{3}{2}} \exp\left(\frac{-a^2}{2r_i^2} | X_i - P_j |^2\right) \quad (5)$$

  Where a is a scale factor for the radii. The value of a=2 is used in all the experiments (for this value, 90% of the electron density lies inside the van-der walls radius of the atom).
- At a given tessellate point P_j, having the surface normal j, the field strength for each donor atom is computed. The direction of each field is given by a unit vector obtained by joining the corresponding atom to P_j. The resultant field at P_j is defined as the vector sum (see Figure 2(c)):

$$\vec{F}(P_j) = \sum_i [f(P_j, A_i) \times (\vec{i} \cdot \vec{j})] \quad (6)$$

In this formulation maximum weight is given to those atoms whose field direction coincide with the surface normal at the specific tessellate point, in computing the resultant field. The acceptor field is analogously defined. Typically Nitrogen or Oxygen atoms with a lone pair of electrons are considered as acceptors. At the end of this stage, the molecule is represented by a set of points at each of which three values corresponding to the geometric shape, donor field, and acceptor field are defined.

4.2. Molecular Similarity

Similarity of two molecules is defined in terms of the similarity of the property distributions. Histogram intersection provides a rapid way to empirically define the similarity of distributions. Furthermore, it is highly efficient to compute and is invariant to translation and rotation. However, since no pose information is stored with the molecular descriptors a direct application of histogram intersection is not possible. We use Figure 3 to show the intuition behind our approach. Two poses of the Asprin molecule with arrows indicating the clusters of tessellate points having the same property value are shown first. The molecule Capsaicin is shown last, having a similar number of points with the same field value. We note that the physical distribution on the sphere of the points having the same property value is distinct and does not change with molecular pose. Furthermore, the pair-wise distances between these points can be used to characterize this distribution.

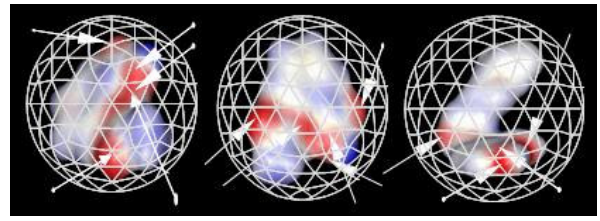

Figure 3: Intuition behind the topological constraint

Based on this our method to obtain the similarity of the distributions can be described as follows:

- For each of the distributions $P_1 \ldots P_K$ used in representing the molecule, define a (fixed) quantization and histogram the distributions. Let H_L denote the histogram corresponding to the distribution P_L, $L \in [1, K]$ and let k denote a specific bin in the histogram.
- Cluster the tessellate points having the same value (falling in the same bin k) by adjacency. Compute the centroid for each cluster.
- Compute the distance (constrained to lie on the surface of the sphere) for all pair of centroids.
- Quantize these distances into bins as follows: {[0,1[, [1,2[, . . .[C/2-1, C/2[}, where C is the circumference of the sphere. Compute the histogram for the distance distribution.
- For two distance histograms D_M and D_I (with the bins indexed by j) the histogram intersections are computed using Eq. (7). The distance histogram intersection defining the topological similarity of values in bin k, denoted by γ_k is defined as the average of the two intersection values from Eq. (7) (to ensure symmetry).

$$H(D_I, D_M) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{M_j}}; H(D_M, D_I) = \frac{\sum_{j=1}^{C/2} \min(D_{I_j}, D_{M_j})}{\sum_{j=1}^{C/2} D_{I_j}} \quad (7)$$

- The topologically constrained histogram intersection value for a given property distribution P_L between two molecules M_1 and M_2 is defined as:

$$H_{TC}(M_1, M_2) = \frac{H(M_1, M_2).\gamma + H(M_2, M_1).\gamma}{2} \quad (8)$$

Where the histogram intersection $H(M_1,M_2)\gamma$ is computed as:

$$H(M_1, M_2) = \frac{\sum_{j=1}^{K} \min(M_{1j}, M_{2j}) \times \gamma_j}{\sum_{j=1}^{K} M_{1j}} \quad (9)$$

- The full histogram intersection $H_{full}(M_1,M_2)$ between two molecules M_1, M_2, is the average over all property distributions of the corresponding topologically constrained histogram intersection values. To account for molecular conformations, we define the similarity of two molecules M_1 and M_2 as the maximum value of the full histogram intersection defined over a set of conformations the molecules can assume:

$$Similarity(M_i, M_j) = \arg\max_{C_i, C_j} [H_{\text{full}}(C_i, C_j)]$$
$$C_i = \{C_i^1, C_i^2, ..., C_i^r\}, C_j = \{C_j^1, C_j^2, ..., C_j^r\} \quad (10)$$

The conformations for a molecule can be generated by finding the local minima of a non-linear function (see [6] for an example), that defines the energy of a molecular structure. This function contains terms capturing the energy due to inter-atomic interactions as well as due to deviations in bond length, bond angles, and tertiary angles. Typically a standard package like CONCORD [24], is used for this purpose.

While developed independently, we note that our idea of constraining the histogram intersection shares similarities with works in image retrieval like [4] and [10] where spatial information has been used in conjunction with color to query 2D image-based data.

5. EXPERIMENTAL RESULTS

To validate the method, two set of experiments were conducted. In the first, the method was tested by computing the similarity for 5000 molecules randomly selected from the MDDR collection [23], which is a commonly used reference in drug discovery and structural biology. It consists of molecules that are either marketed drugs or had reached advanced stages in a drug discovery process. Each molecule from the set of 5000 was used as a query against the rest. The query and model molecules were each represented by 20 conformers, i.e. 400 distinct molecular structures were used per similarity computation. The results for the recognition experiment are presented in Table 1 (second row). The first row of the table shows results obtained on the same set of molecules with ISIS [23], a widely used commercial 2D chemical database. This provided a reference for evaluating the current research. The third row shows results obtained with the proposed approach in a setting where 20 novel (distinct from the model) conformers were used for the query. It may be noted, that for some of the molecules, 20 novel energetically stable conformers could not be obtained. In such cases, the similarity computations involved as many novel conformers as could be derived for each specific case. This is denoted by an asterisk in Table 1. In the third setting, of the 5000 molecules, 4910 were correctly identified.

Method	Data Size	#Conformations	Accuracy
ISIS	5000	none	100%
Proposed	5000	20/20	100%
Proposed	5000	20/20*	98.2%

Table 1: Summary of the similarity experiment

The experiment indicates that the recognition accuracy of the proposed method is comparable with that of ISIS [23] which essentially uses a 2D structural motif-based search algorithm. However, unlike ISIS, the proposed approach can take into account molecular conformations, surface-properties, and superposition-based effects.

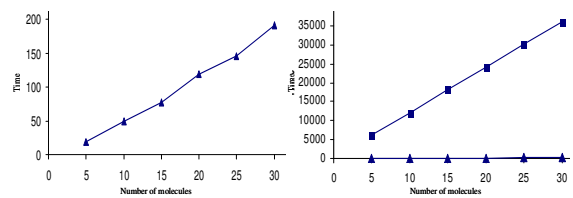

Figure 4: Computational performance of the proposed method (left) and comparison with [8] (right)

In the second experiment, the computational performance of the proposed approach was tested against the system described in [8]. For this 30 pre-selected molecules (see [8] for details) from the MDDR collection were compared against each other, with 20 conformers for the model and one for the query. Both the systems had a 100% recognition rate on this subset of molecules. However, the time requirements were significantly different. A graph plotting the time required for the similarity computation with the proposed technique is shown in Figure 4 (left plot with the data points shown as triangles). Figure 4 (right plot) shows a comparison of the performance with the method outlined in [8] (data points obtained with [8] are shown as squares). On an average, with the proposed technique 120 conformers were processed (descriptor

generation and matching) every second, while with [8], one conformer was processed every two seconds on an SGI Indigo2 machine. Another recent method [12] reports speeds of 2 minutes per molecule (on a SUN Ultra-30).

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we considered the problem of molecular similarity and presented a method motivated by research in computer vision. The recognition accuracy of the proposed method is comparable to both research and commercial systems available in this area. Furthermore, the computational speed of the method, owing to the use of the histogram intersection, is significantly better than currently available 3D molecular matching techniques. This work opens up many directions for further research: Most importantly, its accuracy and speed would allow 3D querying and exploration of chemical space much more easily than currently possible. Preliminary evidence from other ongoing research, leads us to believe that the use of similarity information obtained from the current approach, allows for the construction of better structure-biological activity models (in terms of prediction accuracy) when compared to other 3D matching techniques, while using the same prediction algorithm. With respect to the current line of research, it would be of significant interest to develop error models for recognition accuracy. We also plan to research and develop further the representation approach. The author hopes that this work will help in introducing the computer vision community to some of the significant problems in biology and drug discovery, where vision research can have significant impact.

7. ACKNOWLEDGEMENTS

The author wishes to thank Dr. Ramesh Jain. Many stimulating discussions with him on computer vision played an important role in the development of this work.

8. REFERENCES

[1] Barkat T. and Dean M., *J. of. Comp.-Aided Mol. Design, 4, 107, 1991*

[2] G. Bemis, and I. Kuntz, "A fast and efficient method for 2D and 3D molecular shape description", *J. of Comp. –Aided Mol. Design, 6, 607-628, 1992*

[3] B. Bush and R. Sheridan, "PATTY: A programmable Atom Typer and Languagefor Automatic Classification of Atoms in a Molecular Database", *J. Chem. Inf. Comp Sci., 33, 756-762, 1993*

[4] P. Chang, J. Krumm, "Object Recognition with Color Co-occurrence Histograms", *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 498-504. 1999

[5] R. Cramer, et. al., "Prospoective Identification of Biologically Active Structures by Topomer Shape Similarity Searching", *J. Med. Chem,, 42, pp. 3919-3933, 1999*

[6] P. Finn and L.Kavraki, "Computational Approaches to Drug Design", *Algorithmica, 25, pp 347-371, 1999*

[7] Genome International Sequencing Consortium, "Initial Sequencing and Analysis of the Human Genome", *Nature 409, 860-921, February 2001*

[8] A. Ghuloum, C. Sage, A. Jain, "Molecular Hashkeys: A Novel Method for Molecular Characterization and its Application for PredictingImportant Pharmaceutical Properties of Molecules", *J. Med. Chem, 42, 10, pp 1739-1748, 1999*

[9] M. Hebert, K. Ikeuchi, H. Delingette, "A Spherical Representation for Recognition of Free-Form Surfaces", IEEE Trans. On PAMI, 17, 7, pp 681-689, 1995

[10] J. Huang and R. Zabih, "Combining Color and Spatial Information for Content-based Image Retrieval", *European Conference on Digital Libraries, September 1998*

[11] A. Jain, K. Koile, and D. Chapman, "Compass: Predicting Biological Activity from Molecular Surface Properties. Performance Comparison on a Steroid Benchmark", J. Med. Chem., 37, pp 2315-2327, 1994

[12] P. Labute and C. Williams, "Flexible Alignment of Small Molecules", *J. Med. Chem., 44, 10, pp. 1483-1490, 2001*

[13] I. Lukovits, *J. of Chem. Inf. And Comp. Sci., 31, pp 503, 1991*

[14] M. Miller, "Chemical Database Techniques in Drug Discovery",*Nature Reviews(Drug Discovery) March, 2002*

[15] R. Norel, D. Fischer, H. Wolfson, and R. Nussinov, "Molecular Surface-Recognition by a Computer Vision-Based Technique", *Protein Engineering, 7, 1, pp 39-46, 1994*

[16] M. Papadopoulos, P. Dean, *J. of. Comp.-Aided Mol. Design, 5, 119, 1991*

[17] D. Parsons and J. Canny. "Geometric problems in molecular biology and robotics", Proc. *ISMB* 1994

[18] X. Pennec and N. Ayache, "A Geometric Algorithm to Find Small but Highly Similar 3D Sub-Structures in Proteins", *Bioinformatics, 14, 6, pp 516-522, 1998*

[19] Randic M., *J. Am. Chem. Soc., 97, 6609, 1975*

[20] M. Swain and D. Ballard, "Color Indexing", *Int. J. of Comp. Vision, 7, 1, pp 11-32, 1991*

[21] D. Veber, S. Johnson, H-Y. Cheng, B. Smith, K. Ward, and K. Kopple, "Molecular Properties that Influence the Oral Bioavailability of Drug Candidates", J. Med. Chem, 45, pp 2615-2623, 2002

[22] H. Wiener., *J. of Am. Chem. Soc., 69, pp 17-20, 1947*

[23] http://www.mdli.com

[24] www.tripos.com