Determining Molecular Similarity for Drug Discovery using the Wavelet Riemannian Metric

Elinor Velasquez[1,*], Emmanuel R. Yera[1,*], and Rahul Singh [1,2,*]
Department of Computer Science[1] and Center for Computing in Life Sciences [1, 2]
San Francisco State University, San Francisco, CA 94132
velasque@sfsu.edu, eyera@sfsu.edu, rsingh@cs.sfsu.edu

Abstract

Discerning the similarity between two molecules is a challenging problem in drug discovery as well as in molecular biology. The importance of this problem is due to the fact that the biochemical characteristics of a molecule are closely related to its structure. Therefore molecular similarity is a key notion in investigations targeted at understanding existing molecules as well as in guiding the synthesis of new molecules. Additionally, the notion of molecular similarity plays a central role in structure query-retrieval. This paper presents a Wavelet-based Riemannian metric for determining molecular similarity. The proposed metric extends traditional molecular similarity measures in terms of its ability to capture and compare nonlinear molecular descriptors, thus allowing more accurate characterization of the true nature of the factors involved. Furthermore, owing to its metric properties and wavelet nature, this similarity measure supports highly efficient query-retrieval strategies. To compare graph-based molecular representations using the wavelet-based Riemannian metric, the paper uses a two-phase molecular graph matching strategy. In the first step, an efficient nonlinear graph-matching technique based on the graduated assignment algorithm is used to obtain a preliminary correspondence between molecular graphs in terms of their topological characteristics. Starting from this correspondence, the second stage directly optimizes the proposed metric on arbitrary molecular descriptors using a branch-and-bound search strategy. Various experiments, many in comparative settings, study the retrieval performance of this similarity formulation and underline its efficacy and efficiency.

1. Introduction

Rational drug design is based on the principle that similar molecules have similar biochemical properties. It thus postulates a direct link between the structure of a molecule and its putative biochemical function. A corollary is the notion of *bioisosterism* [3]: similar sub-structural motifs may be interchanged while maintaining similar properties. Together, these ideas lie at the basis of modern pharmaceutical sciences. For instance, in modern drug discovery pipelines, combinatorial chemistry is used to sample the structural space of interest and high throughput screening is employed to test the synthesized molecules for the desired biochemical activity. Following this, in the lead optimization stage, hits are structurally optimized for efficacy, pharmacokinetics, pharmacodynamics, and toxicity. Underlying rational drug design and pharmaceutical techniques lies the problem of determining similarity amongst molecules. Applications of molecular similarity can be classified to have occurred in three directions [3, 22]: biochemical and computational explorations of the molecular structural space, elucidation of structure-property relationships, and structural-based querying. Efficient techniques for determining molecular similarity can form the basis for algorithms exploring the structural space of both small and large molecules [12, 6, 9], structure querying, and structure property relationship modeling [10, 21]. Due to the explosive growth in the sizes of structural databases, the problem of algorithmically determining molecular similarity is rapidly evolving as a critical challenge in modern science.

[*] All authors contributed equally.

Sixth IEEE Symposium on BionInformatics and BioEngineering (BIBE'06)
0-7695-2727-2/06 $20.00 © 2006 IEEE

Four key factors influence the problem of determining molecular similarity: *molecular representation, molecular descriptors, the similarity metric and the matching algorithm.* The research presented in this paper considers determining molecular similarity using graph representations with connectivity-based descriptors and arbitrary physicochemical descriptors. Specifically we investigate this problem in two steps: First, we develop a novel wavelet-based Riemannian metric to capture the notion of molecular similarity. The second step addresses the problem of matching molecular graphs where atoms are vertices and bonds are edges by applying a nonlinear graph-matching technique using the graduated assignment algorithm [11] to the two molecular graphs being compared. The output of this stage is a topologically optimal correspondence between the two molecular graphs. Then, this correspondence is used as an initialization to a branch and bound matching that directly optimizes the proposed metric over a set of molecular descriptors.

The key contribution of this work is to develop a fundamentally rigorous, structurally, bio-chemically and computationally efficient technique to compare molecules by accounting for the non-linear nature of most molecular descriptors. The experimental validations in this paper focus on small drug-like molecules, satisfying Lipinski-like criteria [14] and underline the directly applicability and potential of this research to drug discovery, pharmaceutical sciences, and structural biology.

We begin this paper with an introduction to the problem background and its key characteristics in Section 2. The proposed method is formulated in Section 3. Experimental results are reported in Section 4. The conclusions and directions of future work are presented in Section 5.

2. Problem Background

In their simplest form, molecules can be represented using chemical formulae. However, different structures may yield the same formula (e.g. in the case of isomers), even though they possess dissimilar physical or biochemical properties. Therefore, commonly employed representation frameworks tend to *directly* characterize molecular structure and include: one-dimensional string-based descriptors, such as SMILES or structure keys, obtained by ordered traversal of the molecular graph or representing presence/absence of predefined sub-structural motifs; two-dimensional and three-dimensional graphs characterizing molecular connectivity and inter-atomic distances and three-dimensional surface based representations, such as the Connoly surface. An illustration of these representations, using the Benzene molecule as an example, is presented in Figure 1. The complexity of representations is correlated with their fidelity in describing biochemical characteristics of molecules. It may be noted that molecular graphs, are easier to manually interpret when compared to SMILES strings or surface-based representations, thus the popularity of such representations in medicinal chemistry and in query-retrieval settings.

The idea of molecular descriptors is closely associated with that of molecular representation. Molecular descriptors are properties of the molecule that are of interest. Examples of molecular descriptors include physical-chemical descriptors such as number of rotatable bonds, polar surface area, electronegativity, descriptors of molecular connectivity such as the Wiener number [22], the Randic index [19], structure keys and molecular fingerprints, eigenvalue-based descriptors such as BCUT descriptors [18], molecular moment-based descriptors such as CoMMA [21], and surface and field-based descriptors [22, 10]. Other descriptors include discrete [7] or field-based representations of donor-acceptor atoms and descriptors utilizing the molecular wave and density functions [15]. Certain descriptors are meaningful only in context of specific representations. For example, connectivity indices (Wiener number, Randic index), structure keys, or fingerprints can only be defined on 3D or 2D molecular graphs.

Similarity measures are functions that map pairs of molecular descriptions, a combination of the representation scheme and appropriate descriptors, to real numbers. Typically, dissimilarity measures are used with the value of zero corresponding to identical molecular descriptions. Common dissimilarity measures employed for structural comparison are the Hamming and Euclidean metrics and the Tanimoto measure. Of these, the Tanimoto measure (shown in (Eq. 1)) is the most commonly employed. Here d denotes the distance between molecules Q and R.

$$d = \sum_{1 \le m \le N} x_Q^m x_R^m \Big/ \sum_{1 \le m \le N} (x_Q^m)^2 + \sum_{1 \le m \le N} (x_R^m)^2 - \sum_{1 \le m \le N} x_Q^m x_R^m \qquad (1)$$

These measures have been applied to discrete representations such as bit strings representing structure keys, hashed fingerprints, pharmacophore points or affinity to bind to a panel of receptors [6]. Other similarity measures include root-mean-square error on 3D alignment of structures [13], constrained histogram intersection of property distributions on molecular surfaces [22] and Feature Trees [20] which

lie between bit string descriptors and 3D descriptors, have also been designed. In Feature Trees, the connectivity of hydrophobic fragments and functional groups in a molecule is represented as a tree and similarity is defined through the match of *sub-trees*.

The goal of molecular matching is to obtain the optimal value for the dissimilarity measure for any given pair of molecules. It should be noted that the matching formulation may be algorithmically intractable, for example, in matching molecular graphs or detecting similar sub-graphs. A simplified formulation is that of matching fingerprint or structure keys by directly minimizing the dissimilarity measure. Other formulations include sub-graph and tree matching [*Ibid.*], atom re-labeling to minimize a difference-distance matrix, application of geometric hashing and its variations [16], direct 3D pose optimization [10], and matching molecular surface and field characteristics [10, 13, 22].

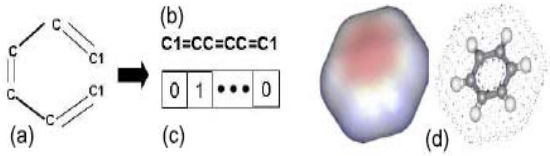


Figure 1 (a) Chemist's representation of benzene, (b) SMILES, (c) bit string, (d) surface-based representation.

3. The Proposed Method

Most molecular descriptors are nonlinear in nature. This is not only true for descriptors based on the wave function, but also for a large class of simplified descriptors used in molecular modeling (such as field-based descriptors) and even highly simplistic descriptors such as molecular mass, which grows nonlinearly with increasing atomic size. For such descriptors, the resulting descriptor space is nonlinear, forms a nonlinear volume in space; the associated space has a curvature. Correctly comparing such descriptors therefore requires a curved-space metric. Commonly used dissimilarity measures such as the Hamming, Euclidean, and Tanimoto measures are, rigorously speaking, inapplicable from this perspective.

3.1 The Wavelet Riemannian Metric

A key characteristic of a nonlinear descriptor space is that it has a curvature. It is therefore more accurate to compare such descriptors using a curved-space metric. For real valued molecular descriptors, the nonlinear descriptor spaces can be described as a Riemannian space. To compare these descriptors, it is natural to use the distance measure that comes with the Riemannian space, namely the Riemannian metric. In the following, we describe the key concepts which help us define a Riemannian space and a metric. For more details on Riemannian spaces we refer the reader to [5].

Definition 1: A topological space, *S*, is a set endowed with a topology, that is, a collection of subsets, so that the empty set and the space *S* belong to the collection of subsets, and the intersection or union of any finite number of subsets belongs to the collection of subsets.

To say that a topological space is differentiable is to imply that the techniques of calculus can be defined on this space. A topological space, *A*, is diffeomorphic to a topological space, *B*, if there is a differentiable mapping from *A* to *B* and a differentiable inverse mapping from *B* to *A*.

Definition 2: A smooth manifold is a topological space that is locally Euclidean; around every point in the manifold there is a neighborhood that is diffeomorphic to a neighborhood in *n*-dimensional real space.

Definition 3: A smooth manifold, *M*, is called a Riemannian manifold if it is equipped with a symmetric positive definite 2-tensor, *g*, namely a Riemannian metric. Given the set of g_{mn} are the components of a matrix representation of *g*, let *d* be the arc length on M. A Riemannian metric is defined as $d = \sum_{1 \le m,n \le N} g_{mn} dx^m dx^n$. The matrix *g* is a positive definite, symmetric matrix since it is the matrix representation of a symmetric positive definite 2-tensor.

The simplest type of metric which subsumes most known similarity metrics is a Riemannian metric: if we set the matrix *g* to be the identity matrix, we arrive at the Euclidean metric. There are an infinite number of possible Riemannian metrics since the matrix coefficents g_{mn} are differentiable functions on *n*-dimensional real space; formally, the matrix *g*, with matrix components g_{mn}, is defined to equal the inner product of two basis vectors from the tangent space of the Riemannian space. Therefore we need to set some constraints to compute the Riemannian metric that will serve as our similarity metric. We especially need to model the metric using functions that can handle nonlinearity and can support efficient computations. A class of functions that have these properties are wavelets [23].

3.2 Derivation of the Metric

We use the Morlet mother wavelet to approximate the g_{mn} of the Riemannian metric:

$\psi(x) = \cos(x)\, e^{-x^2/2\sigma^2} / \sigma\sqrt{2\pi}$. Setting $g_{mn} = \int \psi(mt)\psi(nt)dt$ gives that g_{mn} is a function of the error function:

$$Erf(z) = 2\int_0^z e^{-t^2} dt \,/\, \sqrt{\pi}. \quad (2)$$

Taking a first order approximation of the *Erf (z)* yields the proposed metric. The matrix g is constrained to be positive definite, thus we take the absolute value of one term: If N is the number of descriptors, m and n the sub-indices of g_{mn} , x_Q ,x_R the descriptors of molecules Q and R and σ a parameter, then

$$B = |\sigma(n-m)| \Big/ \sqrt{2(m^2+n^2)}, c = 2mn\sigma^2 / m^2 + n^2$$

$$A = \left(m^2+n^2\right)\sum_{1\le m\le N}\left(x_Q^m - x_R^m\right)^t\left(x_Q^m - x_R^m\right) \Big/ \sigma\sqrt{2\left(m^2+n^2\right)},$$

$$a = e^{-2mn(\sigma^2+m^2+n^2)/2(m^2+n^2)} \Big/ 4\sigma\pi\sqrt{2(m^2+n^2)};$$

$$g_{mn} = a\left(A\left(2+e^c\right) + e^{c-A^2+B}\left(A\cos AB/\pi - B\sin AB/\pi\right)\right).$$

Linearizing *x(t),* using a step size of 1 for *dt* after discretization (and using the value of σ =1 throughout), the distance between molecule Q and molecule R becomes: $d^2 = \sum_{1\le m,n\le N} g_{mn}$. The proposed distance function is a wavelet function and a metric, satisfying the four properties of a metric, namely: $d(Q,R)\ge 0$ for every choice of Q and R, $d(Q,R) = 0$, iff $Q = R$, $d(Q,R) = d(R,Q)$ (symmetry) and for an arbitrary molecule X, $d(Q,R) \le d(Q,X)+d(X,R)$(triangle inequality).

The two properties of the metric that are of special interest given the growth in sizes of molecular databases are: (1) The triangle inequality, which can be used to reduce the number of explicit comparisons during query-retrieval (See Section 4.4) and (2) The wavelet nature which allows using a small number of coefficients to finely approximate the similarity between molecules (See Section 4.5).

3.3 Correspondences between Molecular Graphs

Given a similarity metric, determining molecular similarity involves obtaining the correspondence between two molecules. In the case of graph-based representations, this requires graph matching. Graph matching techniques can be classified as those that construct a state-space and attempt to search it using branch and bound techniques and non-linear optimization methods such as relaxation labeling, linear programming, graduated assignment [11], SVD-based eigen-decomposition, and thin-plate spline-based point matching. Heterogeneity of atoms and bond types, and the complexity of molecular descriptors, however complicates a direct application of these techniques to molecular matching.

A molecular similarity formulation needs to deal with whole molecular or sub-structural querying. These query modalities capture different underlying semantics; whole-molecular querying is often exploratory and used for generating hypotheses while sub-structural querying requires users to have a clear idea of the structures which need to be retrieved [22].

It should also be noted that the ability to determine similarity of molecules from the similarity of its components (sub-structures, atoms) is critical because it provides cues on how to leverage *bioisosterism,* e.g., to optimize molecular structure-activity relationships. Most commonly used formulations such as bit string comparisons are not decomposable beyond a certain point. Neither structure keys nor fingerprints can be reliably used to discern atom or bond-level variations.

Three types of matching formulations may be conceived: In *exact matching* there is, up to Euclidean transformations, a one-to-one onto correspondence between two molecular graphs. In *substructure matching* a common substructure between two molecules needs to be found. Finally, determining all other forms of similarity falls under the challenge of *inexact matching*.

Graph matching is considered to be of high complexity owing to the combinatorial nature of the problem. Exact graph matching has been shown to be of polynomial time, while sub-graph and inexact graph matching is known to be NP-complete [4]. One way to look at graph matching is as a solution to the registration problem between two sets of points (features). In such problems, due to factors such as noise or inherent nature of the data, it is possible that the feature points can not be exactly matched or that points may exist in one set that have no corresponding points in the other set. A relatively recent result in the area is a nonlinear optimization technique known as the graduated assignment algorithm (GAA) [11] where an estimation of correspondence between the two sets is obtained to match them. In the context of matching molecular graphs, the GAA can be understood as follows: The correspondence is viewed as a linear assignment problem [17]. For two molecular graphs GAA attempts to find a correspondence matrix M such that a quadratic energy function, explicitly dependent on M, is minimized. This energy function is, with *V(1)* and *V(2)* vertex sets of Graph 1 and 2, respectively,

$$E(M) = \frac{-1}{2}\sum_{a=1}^{V(1)}\sum_{i=1}^{V(2)}\sum_{b=1}^{V(1)}\sum_{j=1}^{V(2)} M_{ai}M_{bj}C_{aibj} \quad \text{with}$$

$$\sum_{i=1}^{N+1} M_{ai} = \sum_{a=1}^{K+1} M_{ai} = 1. \quad (3)$$

The GAA computes a matrix which assigns a one-to-one correspondence between the two molecular graphs. The columns represent the vertices of the first graph and the rows represent the vertices of the second graph. The correspondence matrix allows for partial matches between the two molecular graphs. The order of computational complexity is *O(np),* in which *n* and *p* are the vertex size of each of the two graphs, respectively. Our investigations show that the GAA gives good correspondences for exact graph matching. However, for substructure and inexact matching, GAA is not as efficient; it considers points which are non-corresponding as outliers and discards them. In the given context, the sub-optimality of the GAA-based correspondences is also due to the fact that GAA works with the topology of the molecular graph only and does not take into account any chemically relevant characteristics of the molecules.

We treat the correspondences obtained from GAA as an initial approximation to the desired correspondence and use a branch and bound strategy (BBA) as the second step in the matching process. The BBA optimizes the wavelet metric through a breadth-first traversal-based branch and bound strategy as follows. The GAA match matrix dictates that a given vertex on a graph is mapped to a corresponding one on a second graph. The key step lies in checking that the neighborhood around the vertex is appropriately mapped to the neighborhood of the corresponding vertex. Scoring the correspondences using the wavelet Riemannian metric, allows incorporating the relevant descriptors in this process. The connectivity and atom/bond typing of the first molecule is preserved (to the extent possible) in the mapping to the second graph. For each vertex, the wavelet metric is computed to keep a running cost of the graph matching. When the GAA algorithm fails to give an optimal vertex mapping, the cost function or similarity metric is used to give the local maximum, i.e. the best fit for the vertex image. Complexity for the BBA algorithm runs between *O(N),* which is the best case, and *O(N!),* which is the worse case. Our studies show that the worse case is rarely achieved.

4. Experimental Evaluations

A series of experiments were designed to study the proposed metric and similarity formulation. Two data sets consisting of 10,000 and 1,000 molecules each were used in this study. These molecules were randomly selected from the ChemDB database at UC-Irvine [24]. The only constraint observed in selecting the molecules was that the structures be connected. The larger 10,000 molecule dataset was used for large-scale experiments, while the 1,000 molecule dataset was used in experiments that involved exhaustive expert validation. The smaller dataset consisted of 636 unique molecules, with the other structures being conformers of these molecules. This composition allowed us to study the efficacy of the method in dealing with conformations. The descriptors used in the experiment were mass, number of bonds, electronegativity and variance of local pairwise distances. These descriptors represented a spectrum of descriptor types capturing steric as well as biochemically relevant molecular features. The experimental design studied the method through various perspectives including performance in a query-retrieval setting using both whole-molecular and sub-structural search, comparisons of the retrieved results using the wavelet metric with those using Tanimoto measure (keeping the same matching algorithm), ability to group conformers, reducing the search space by using the metric character of the measure, and capability of the proposed metric to capture similarity at increased resolution by using a larger number of wavelet coefficients.

Table 1.
Percentage query is retrieved first in rankings

Database size	Number of queries	% Accuracy
1,000	1,000	100%
10,000	200	100%

4.1 Accuracy in query-retrieval settings

In the first part of this experiment, the efficacy of the method for whole molecule querying was tested on both the 1,000 and 10,000 molecule datasets. For the former, each of the 1,000 molecules was successively used as a query against the rest of the molecules in this set. Prior to the query, each query molecule underwent a random rotation. The accuracy of the method was determined by its ability to retrieve the query molecule as the top ranked hit (lowest dissimilarity) from the database. A smaller set of 200 randomly selected queries was also run on the larger, 10,000 molecule dataset in this setting. These results are presented in Table 1 and experimentally illustrate the invariance of the technique to Euclidean transformations.

In the second part of this experiment, the ability of the method to handle sub-structure queries was tested using precision and recall values determined through a manual analysis of the data. This experiment was run on the 1000 molecule dataset and manually analyzed to determine the precision and recall. A subset of these results on the 1000 molecule dataset (showing ten of the substructures used as queries) was presented in Figure 2.

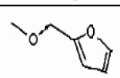
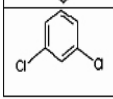
Substructure	Precision	Recall	Substructure	Precision	Recall
	100%	100%		100%	97.6%
	100%	100%		100%	100%
	100%	100%		100%	100%
	100%	100%		100%	100%
	100%	100%		100%	100%

Figure 2: Examples of the substructure search with corresponding precision-recall values

4.2 Comparison with the Tanimoto Measure

Queries were conducted against the 10,000 molecular dataset with the proposed matching technique used with the wavelet Riemannian metric and the Tanimoto measure. Fixing the matching technique allowed us to factor it out and compare the proposed metric and the Tanimoto measure in terms of the quality of retrieval. Analyzing similarity of derived structures was context dependent and thus ultimately subjective. Our study considered the top ranked molecules retrieved for a query and analyzed them in terms of how well key structural patterns such as scaffolds and multi-rings were preserved, for each of the similarity measures. Some examples are presented in Figure 3. The triple-ring motif of the query was not preserved in the third ranked retrieval using the Tanimoto; it was retained in the second, third, forth, and fifth ranked retrievals using the wavelet metric. In the next example, the Tanimoto comparison retrieved a two-ringed structure as the second-most similar hit, while the wavelet measure retrieved a structure much more similar to the query. A general analysis of 100 random queries, showed that the proposed metric retrieved more similar molecules than the Tanimoto measure in 48% of the cases, was indistinguishable in 35% of the cases and retrieved worse molecules than the Tanimoto measure 17% of the time. Even in these 17% of cases, the wavelet metric resulted in the most similar molecule to the query (excluding itself) being retrieved within the top ten results.

4.3 Ability to Group Conformers

We studied the ability of the proposed approach to group conformers. For our descriptors (topological and local steric shape), conformers should be grouped together. We compared the results with those obtained using the Tanimoto measure. The top plot in Figure 4 presents these results. As expected, the conformers are represented by the flat sections of the plot. Two observations can be made: First, the wavelet-based approach distinguished between the top hits better than Tanimoto-based search. The distance between the top two hits was significantly more pronounced with our approach. Second, the conformers at the plateaus labeled "*b*" and "*c*" were interchanged when comparing the Tanimoto measure to the wavelet metric. Inspection of these structures indicated that the wavelet method retained structural similarity better.

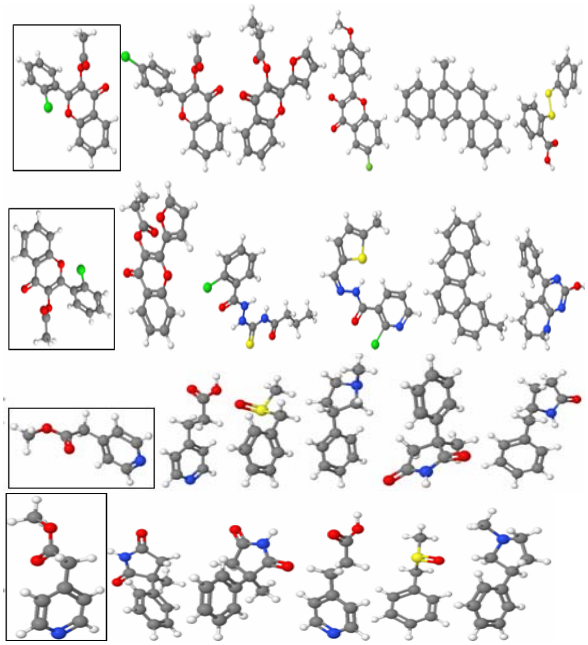


Figure 3: First and third rows: The query molecule and first through fifth ranked retrievals using the wavelet metric; Second and fourth rows: The corresponding retrievals using the Tanimoto measure. In all cases the query molecule is shown in a box.

4.4 Reducing the Search Space

For a similarity measure that has metric properties, the triangle inequality can be used to avoid an exhaustive search of the database in retrieving the molecule most similar to the query. The basic idea lies in determining how molecules in the database are related to a predefined reference molecule. If the similarity between the query and the reference molecule can be computed, then the molecules in the database that are highly dissimilar from the query can be excluded without resorting to costly on-line computations. Let Q, be the query molecule, R, the reference molecule, and X, an arbitrary database molecule. A "joint cutoff criterion" can be derived

from the triangle inequality such that for any X, if $|d(X,R)-d(Q,R)|>\xi$, with $\xi = d(Q,B)$ the minimum distance between Q and the closest molecule B, found so far in the search, then X can be discarded from the similarity search (as B is more similar to the query than X). Selecting the proper reference molecule plays an important role. While this issue is outside the scope of the current paper, we show in Table 2, the reduction in search obtained by using a random molecule as a reference. It may be noted that the reduction in the search space varied between 1% and 99% without influencing the correctness of the top retrieved molecule. Typically, for diverse data sets like the one used in this study, multiple reference points are used.

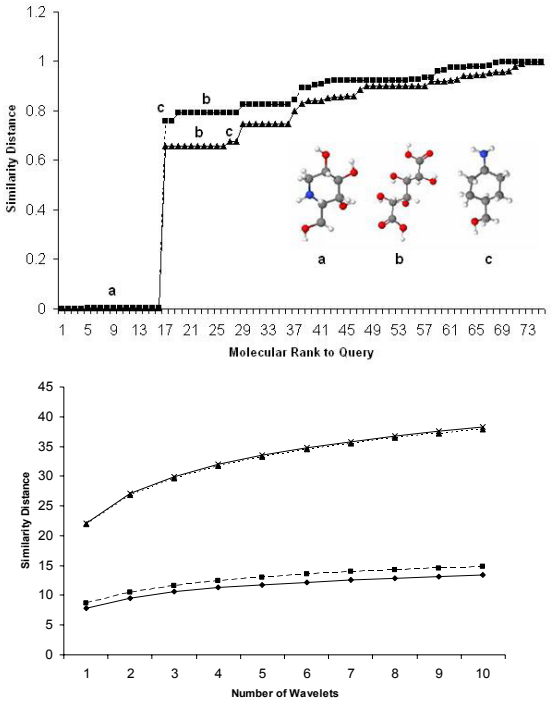


Figure 4: Top graph: grouping conformers. Lower curve: Tanimoto measure (triangles); Upper curve: Proposed metric (squares). Bottom graph: Wavelet metric showing variations in similarity scores for the first ranked molecule (diamond), second ranked molecule (square), and the third ranked molecule (triangle) as the number of wavelet coefficients are increased.

4.5 Capturing Similarity at Increasing Resolutions

By increasing the number of wavelets used in the estimation of the Riemannian metric, one can increase the "resolution" at which similarity between two molecules is determined. We investigated this property as follows: For a query, the top 1000 retrievals were ranked with respect to their distance from the query. We then successively added wavelets and observed the behavior of similarity scores. For instance, the measure using five wavelets is defined as:

$$\sum_{1\le m,n\le N} g_{mn}(x)+\sum_{1\le m,n\le N} g_{mn}(x/2)+\sum_{1\le m,n\le N} g_{mn}(x/3)+ \sum_{1\le m,n\le N} g_{mn}(x/4)+\sum_{1\le m,n\le N} g_{mn}(x/5) \quad (4)$$

These results were presented in Figure 4, where the distance of the query from the first three molecules was plotted as the number of wavelets increased. As the dissimilarity between molecules increased, using larger number of wavelets led to better distinction (larger distances) than using a single wavelet. This difference started to saturate after 10 wavelets. However, for highly similar molecules, such an effect was less pronounced, thus indicating that increasing the number of wavelets indeed did add to the ability to discern molecules rather than simply adding to the distance.

Table 2.
Reduction in search space by employing the triangle inequality for randomly selected reference molecule.

Query	Database size	Number of comparisons	% of Molecules Excluded from Search
1	1000	428	57.2%
2	1000	388	61.2%
3	1000	994	6.0%
4	1000	998	2.0%
5	1000	994	6.0%
6	1000	366	63.4%
7	1000	131	86.9%
8	1000	206	79.4%
9	1000	7	99.3%
10	1000	238	76.2%
Average	1000	475	53.8%

5. Conclusions

This paper introduces a novel wavelet-based Riemannian metric for determining molecular similarity. Utilizing the metric, the problem of defining similarity between molecules is considered on graph-based molecular representations. The wavelet similarity measure allows comparison of molecular properties by taking into account their nonlinear nature. This leads to more accurate interpretation of the physics of the problem. From a theoretical perspective, the wavelet metric subsumes popular similarity measures like the Hamming and Euclidean distances. Additionally, it satisfies metric properties. From an algorithmic perspective, its wavelet and metric nature can be utilized to reduce the number of necessary comparisons in a query-retrieval setting. This underlines its potential for use in structural querying of large molecular databases. Finally, from a biochemical perspective, experiments indicate that the proposed

metric is competitive with the established Tanimoto similarity measure and superior to it in terms of its descriptive and algorithmic advantages.

A graph matching formulation, combining a graduated assignment-based non-linear optimization step with a branch and bound phase is used to compute similarity between molecules using the wavelet-based Riemannian metric. Experiments indicate that this formulation can be used for highly-accurate query-retrieval. The theoretical and experimental validation presented in the paper underlines the fundamental nature of this contribution and its potential to find broad applicability in a number of scientific investigations where molecular similarity plays a crucial role. Future work involves investigations surrounding the reference molecules when employing the triangle inequality.

6. References

[1] Jürgen Bajorath, Ed., Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery, Humana Press (Totowa, NJ: 2004).

[2] J. Barnard, G. Downs, and P. Willett, "Descriptor-Based Similarity Measures for Screening Chemical Databases", pp. 59-80 in Virtual Screening for Bio-Active Molecules, Methods and Principles in Medicinal Chemistry, Vol. 10, Eds H-J Bohm and G. Schneider, Wiley-VCH, 2000

[3] Andreas Bender and Robert C. Glen, Molecular similarity: A key technique in molecular informatics, Org. Biomol. Chem. (Vol. 2, No. 22), 2004, pp. 3204—3218

[4] Endika Bengoetxea, Inexact graph matching using Estimation of distribution algorithms, PhD Thesis, 2002, Ecole Nationale Supérieure des Télécommunications (Paris)

[5] Boothby, William M., An introduction to differentiable manifolds and Riemannian geometry, Academic Press (New York: 1975)

[6] H. Briem and I. Kuntz, "Molecular Similarity Based on Dock-Generated Fingerprints", Journal of Medicinal Chemistry, Vol. 39, 1996, pp. 3401—3408

[7] 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, 1993, pp. 756—762

[8] H. C. Kolb, "Click Chemistry: Diverse Chemical Function from a Few Good Reactions", Angew. Chem. Int. Ed. Vol. 40, No. 11, 2001, pp. 2004—2021

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

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

[11] S. Gold, and Rangarajan, A., A graduated assignment algorithm for graph matching. IEEE Trans. Patt. Anal. Mach. Intell. 18 (4), pp. 377—388

[12] Holm L., and Sander C., "Mapping the protein universe", Science, 273, 1996, pp.595—603

[13] 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, 1994, pp 2315—2327

[14] CA Lipinski, Adv. Drug Del. Rev. 1997, 23, 3

[15] Nina Nikolova and Joanna Jaworska, Approaches to measure chemical similarity – a review, QSAR Comb. Sci., 22, 2003, pp. 1006—1026

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

[17] Christos H. Papadimitriou and Kenneth Steiglitz, Combinatorial Optimization: Algorithms and Complexity, Prentice-Hall, Inc. (Englewood Cliffs, NJ: 1982).

[18] R.S. Pearlman, K.M. Smith. Metric validation and the receptor-relevant subspace concept, J. Chem. Inf. Comput. Sci., 39, 1999, pp. 28—35

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

[20] Matthias Rarey and J. Scott Dixon, Feature Trees: A new similarity measure based on tree matching, J. Comp.-Aided Mol. Design (Vol. 12) 1998, pp. 471—490

[21] B.D. Silverman, D.E. Platt, Comparative molecular moment analysis (CoMMA): 3D-QSAR without molecular superposition, J. Med. Chem. 1996, 39, pp. 2129—2140

[22] R. Singh, "Reasoning about Molecular Similarity and Properties", Proc. IEEE Proc. IEEE Computational Systems Bioinformatics Conference (CSB), 2004

[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, 2002, pp. 2615—2623

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

[23] David F. Walnut, An Introduction to Wavelet Analysis, Birkhauser (Boston: 2002).

[24] UC Irvine ChemDB, http://cdb.ics.uci.edu/CHEM/Web