*Journal of Mathematical Chemistry Vol. 35, No. 3, March 2004 (© 2004)*

# From shape similarity to shape complementarity: toward a docking theory

Michel Petitjean

*ITODYS (CNRS, UMR 7086), 1 rue Guy de la Brosse, 75005 Paris, France*
E-mail: petitjean@itodys.jussieu.fr

Received 26 September 2003; revised 9 March 2004

Formal relations between similarity and docking are analyzed, and a general docking theory is proposed for colored mixtures of multivariate distributions. $X$ and $Y$ being two colored mixtures with given distributions, their shape complementarity coefficient is defined as the lower bound of the variance of $(X - Y)' \cdot (X - Y)$, taken over the set of joint distributions of $X$ and $Y$. The docking is performed via minimization of the shape complementarity coefficient for all translations and rotations of the mixtures. The properties of the docking criterion are derived, and are shown to satisfy the practical requirements encountered in molecular shape analysis.

**KEY WORDS:** colored mixture, Wasserstein distance, similarity, docking, shape complementarity coefficient

**AMS subject classification:** primary 60E15; secondary 51M16, 92E10

## 1. Introduction

Molecular shape analysis is usually viewed, either from shape similarity, or from shape complementarity. A vast literature is available on similarity (e.g. see [1–4] and references cited). Although similarity is deduced from some virtual superposition of molecules by means of the computer, shape complementarity is related to real situations, such as in the key–lock model of enzyme–ligand complexes. Searching for shape complementarity by computer is usually called docking, and an abundant literature is also available (e.g. see [5] and references cited). Another difference between similarity and docking procedures is that superposing molecules is viewed as a global or local volume overlap operation, although docking is mostly viewed as a local surface interaction.

The superposition and the docking of two molecules by computer have a common feature: a scoring function is selected as an objective criterion of similarity or complementarity, and this criterion is optimized by rotation and translation of one of the molecules. Thus, the basic requirement of most similarity and docking procedures is to define a suitable criterion measuring the degree of similarity or complementarity of the shapes.



0259-9791/04/0300-0147/0 

The scope of this paper is to show that both similarity and docking may be performed by computer via a common general molecular model introduced recently, called the colored mixture model [6], and that a docking criterion is available for this model, thus working both for discrete and continuous sets, with or without charges. Although a similarity criterion having this properties is known [6], it seems that no docking criterion having them was previously reported in the literature.

## 2. Theory

### 2.1. Colored mixtures

We define a probability space $(C, A, P)$, where $C$ is a non-empty set called the space of colors, $A$ is a $\sigma$-algebra defined on $C$, and $P$ is a probability measure. In its simplest form, $C$ is a finite set of non-ordered elements (the colors), without any particular structure. $C$ may have an infinite cardinality, e.g. it may be isomorphic to $R$. Then, we define a mapping $\Phi$ from $C$ on the space of the probability distributions on $(R^d, B)$, where $B$ is the Borel $\sigma$-algebra of $R^d$. In other words, to each color $c$ is associated a $d$-variate distribution $\tilde{P}_c = \Phi(c)$.

The main idea behind the colored mixture concept, is a two-step process: get a color from the $P$ distribution, then get a $d$-tuple from the distribution $\tilde{P}_c = \Phi(c)$. We consider the measurable space $(C \times R^d, A \otimes B)$, and we consider a random variable $(K, X)$ on this compound space. The value of the distribution function of $\tilde{P}_c$ at $x$ is a conditional probability noted $\tilde{F}(x|c)$.

$X$ is called a colored mixture when its distribution function $F$ satisfies to the Bayesian expression:

$$F(x) = \int_{c \in C} \tilde{F}(x|c) \cdot P(dc). \tag{1}$$

The differences between $X$ and a random variable on $R^d$ having a mixture distribution, in its usual sense [7], are that, for the latter, there is no space of colors, $c$ is a finite integer index, and a finite summation is used rather than a Lebesgue–Stieltjes integral. When $K$ is almost surely equal to some constant (e.g. when $C$ contains only one element), there is no essential difference between $X$ and an ordinary random vector.

### 2.2. The colored mixture model

We consider the couple of random variables $((K_1, X_1), (K_2, X_2))$, where $X_1$ and $X_2$ are colored mixtures. The joint distribution of $(K_1, K_2)$ is $P_{12}$. We have also a couple of mappings $(\Phi_1, \Phi_2)$ and then, for each couple of colors

$(c_1, c_2)$, we have a couple of $d$-variate distributions $(\Phi_1(c_1), \Phi_2(c_2)) = (\tilde{P}_{c_1}, \tilde{P}_{c_2})$. This latter couple has a joint distribution function, and its value at $(x_1, x_2)$ is a conditional probability noted $\tilde{W}(x_1, x_2|c_1, c_2)$. The joint distribution function $W(x_1, x_2)$ of $(X_1, X_2)$ is got by integration:

$$W(x_1, x_2) = \int_{c_1 \in C} \int_{c_2 \in C} \tilde{W}(x_1, x_2|c_1, c_2) \cdot P_{12}(dc_1, dc_2). \tag{2}$$

Now, the colored mixture model is defined from the following supplementary assumption: $K_1$ and $K_2$ are almost surely equal. It means that $K_1$ and $K_2$ are identically distributed and fully correlated, as expressed in equation (3), where $\delta$ is the Dirac delta function:

$$P_{12}(dc_1, dc_2) = P(dc_1) \cdot \delta_{[c_2=c_1]} dc_2. \tag{3}$$

Using equation (3) for integrating (2), the joint distribution function of $(X_1, X_2)$ is:

$$W(x_1, x_2) = \int_{c \in C} \tilde{W}(x_1, x_2|c) \cdot P(dc). \tag{4}$$

In general, the colored mixtures $X_1$ and $X_2$ cannot be independent.

For example, consider a set $C$ containing two colors $c$ and $c'$ such that $P(c) = P(c') = 1/2$, and the colored mixtures $X_1$ and $X_2$ such that: $\{\text{Prob}(X_1 = a|c) = 1;\ \text{Prob}(X_1 = b|c') = 1\}$ and $\{\text{Prob}(X_2 = a|c) = 1;\ \text{Prob}(X_2 = b|c') = 1\}$, where $a$ and $b$ are some distinct constants in $R^d$. Clearly, the marginals $X_1$ and $X_2$ are such that $\text{Prob}(X_1 = a) = 1/2$, $\text{Prob}(X_1 = b) = 1/2$, $\text{Prob}(X_2 = a) = 1/2$, $\text{Prob}(X_2 = b) = 1/2$, although their joint distribution is such that $\text{Prob}(X_1 = a, X_2 = a) = 1/2$, $\text{Prob}(X_1 = b, X_2 = b) = 1/2$. Since $\text{Prob}(X_1 = a, X_2 = a)$ differs from $\text{Prob}(X_1 = a) \cdot \text{Prob}(X_2 = a)$, $X_1$ and $X_2$ are dependent.

The correlation in the space of colors induces a correlation between the random vectors. When $C$ contains only one element, there is no constraint on the joint distribution of $X_1$ and $X_2$.

The colored mixture model will be further assumed.

### 2.3. *The finite discrete colored mixture model*

The finite discrete colored mixture model is a colored mixture model satisfying to the conditions (a) to (e), given below.

(a) The mixing distribution of the colors is discrete and finite.

(b) The mixed distributions are discrete and finite.

Conditions (a) and (b) means that $X_1$ and $X_2$ are finite colored mixtures of finite discrete distributions. The joint probability of the couple $(X_1, X_2)$ is deduced from equation (4):

$$\text{Prob}(x_1, x_2) = \sum_c \tilde{W}_c(x_1, x_2) \cdot P(c). \tag{5}$$

In equation (5), $c$, $x_1$ and $x_2$, take a finite number of values. Thus, for each $c$ value, the probabilities $\tilde{W}_c(x_1, x_2)$ are the elements of a rectangular bistochastic matrix. Then, the probabilities $\text{Prob}(x_1, x_2)$ are the elements of a rectangular bistochastic matrix $Q$, which has a block diagonal structure: each rectangular block is biunivocally associated to a color, and all elements of $Q$ outside the blocks are null.

The number of colors is $k$, the numbers of lines of the blocks are $m_1, m_2, \ldots, m_k$, and the numbers of columns of the blocks are $n_1, n_2, \ldots, n_k$. We set:

$$m = \sum_{c=1}^{c=k} m_c, \quad \text{and} \quad n = \sum_{c=1}^{c=k} n_c.$$

We assume that the conditions (a) and (b) are satisfied and we add three supplementary conditions:

(c) For each color, the two marginals of $\tilde{W}_c$ are distributed on an equal number of values.

(d) For each color, the two marginals of $\tilde{W}_c$ are uniformly distributed.

(e) The marginals $X_1$ and $X_2$ are uniformly distributed.

Condition (c) means that $Q$ has square blocks, and $m_c = n_c$ for $c = 1$ to $c = k$. It follows that $Q$ is a block diagonal square matrix.

Condition (d) means that each block of $Q$ has all its line sums equal to $P(c)/m_c$, and all its column sums equal to $P(c)/n_c$, $c$ being the color associated to the block.

Now, all conditions (a) to (e) are assumed to be satisfied. The products $P(c)/n_c$ and $P(c)/m_c$ are equal to $1/n$ and do not depend on $c$, because each of the $n$ line sums and $n$ column sums of $Q$ is equal to $1/n$, $n$ being the number of lines (or columns) of $Q$. In other words, for each of the colored mixtures $X_1$ and $X_2$, the probability to get a color is proportional to the number of points of the distribution attached to this color: $P(c) = n_c/n$.

Assuming the finite discrete colored mixture model, we consider the set of all values taken by the elements $\{nQ_{i_1 i_2}\}$ of the matrix $nQ$. These elements are all non-negative, some of them being always null, each line of $nQ$ sums to 1,

and each column of $nQ$ sums to 1. Thus $\{nQ_{i_1 i_2}\}$ is a closed bounded convex set, and its extreme values are such that each square block of $nQ$ is a permutation matrix, and $nQ$ itself is a permutation matrix.

Two particular situations of the finite discrete colored mixture model are of special interest. They are better described using a non-probabilistic formulation:

- Setting $k = 1$: there are two sets of $n$ points, the $2n$ points having all the same color, or simply having no color at all. The extreme values of $\{nQ_{i_1 i_2}\}$ are the $n!$ permutation matrices. This is a non-discernible particles model.
- Setting $k = n$: there are $n$ colors. For each of the two sets of $n$ points, each point is biunivocally associated to a color, and thus there is a one-to-one pairwise correspondence between the two sets of $n$ points. Moreover, the $n$ blocks having only one element, $nQ$ is always the identity matrix. This is a discernible particles model.

## 3. Measuring similarity with the Wasserstein distance

As mentioned, the similarity concept is related to the distance between elements of a metric space. There are many probability metrics measuring the distance between two distributions [8]. Among them, the Wasserstein distance $D$ has received much attention, partly due to its connections with the Monge–Kantorovitch transportation problem [9]. When $X_1$ and $X_2$ are two random vectors, $D$ is expressed from the lower bound of a moment (i.e. an expectation), taken over the set of all the joint distributions of the couple $(X_1, X_2)$:

$$D^2 = \mathrm{Inf}_{\{W\}} E[(X_1 - X_2)' \cdot (X_1 - X_2)]. \tag{6}$$

Both $X_1$ and $X_2$ are assumed to have a finite inertia, in order to ensure the existence of the expectation in equation (6). Now, assuming that we are in the context of the colored mixture model, the set of the joint distributions of the couple $(X_1, X_2)$ is a non-empty subset of the previous one, because adding colors introduces constraints on this set. Thus, we still use equation (6) to define the distance between colored mixtures. The metric properties of the Wasserstein distance are obviously kept.

In many practical situations, rotations and translations of one of the colored mixtures are considered. We assume that $X_1$ is fixed and that $X_2$ is submitted to a rotation $R$ and a translation $t$. Now, we compute the similarity from the minimized Wasserstein distance over the set of rotations and translations. This minimized distance is a similarity coefficient $S_{12}$ called here the intrinsic Wasserstein distance between $X_1$ and $X_2$:

$$S_{12} = \mathrm{Min}_{\{R,t\}} D. \tag{7}$$

The optimal translation $t$ is got when $X_1$ and $X_2$ have the same expectation, and the optimal rotation is known for $d = 2$ and $d = 3$ [6]. However, the optimal joint distribution in equation (6) is not ensured to exist [8,9].

When $X_2$ is distributed as a mirror image of $X_1$, the squared intrinsic distance, normalized to the common inertia $T$ of $X_1$ and $X_2$, is proportional to the chiral index $\chi$ [6]:

$$\chi = d \cdot S_{12}^2 / 4T. \tag{8}$$

Now $X_1$ and $X_2$ are not necessarily distributed as mirror images. We consider finite colored mixtures of finite discrete distributions, i.e. the conditions (a) and (b) in Section 2.3 are satisfied. Looking at equation (5) shows that minimizing any absolute moment of the couple $(X_1, X_2)$ leads to minimizing a linear function under linear constraints, for which the probabilities $P(c)$ are fixed parameters, and the unknown quantities are either the probabilities $\tilde{W}_c(x_1, x_2)$, or the non-null elements of the rectangular bistochastic matrix $Q$. Clearly, the set of linear constraints is a closed bounded convex set, and the minimized moment is reached at least on one of the extremal points of the convex polytope of the constraints [10].

We consider now the finite discrete colored model, i.e. the five conditions (a) to (e) in section 2.3, are satisfied. Discarding rotations and translations, computing $S_{12}$ or minimizing any absolute moment leads to the enumeration of the permutation matrices $nQ$, $n$ being the number of points attached to the distribution of $X_1$ or $X_2$. From equations (6) and (7), we have:

$$S_{12}^2 = \mathrm{Min}_{\{R,t,Q\}}[\sum_{i_1=1}^{i_1=n} \sum_{i_2=1}^{i_2=n} (x_{1_{i_1}} - x_{2_{i_2}})' \cdot (x_{1_{i_1}} - x_{2_{i_2}}) \cdot n Q_{i_1 i_2} / n]. \tag{9}$$

Since $Q$ is a permutation matrix, $n^2 - n$ elements $Q_{i_1 i_2}$ are null, and the others are equal to 1. It follows that the index $i_2$ is biunivocally associated to $i_1$ via the permutation $q$ associated with the permutation matrix $Q$, i.e. $i_2 = q(i_1)$ and equation (9) is rewritten:

$$S_{12}^2 = \mathrm{Min}_{\{R,t,q\}}[\sum_{i=1}^{i=n} (x_{1_i} - x_{2_{q(i)}})' \cdot (x_{1_i} - x_{2_{q(i)}})]/n. \tag{10}$$

The quantity $S_{12}$ in equation (10) is just the well-known Root Mean Square (RMS) criterion of spatial alignment, which is widely used in chemistry and biochemistry (see [11] and references cited). It is also known in data analysis as the pure rotation Procrustes criterion for superposing optimally two groups of points: see appendix A in [6] for various Procrustes methods and their analytical solution.

$S_{12}^2$ is the mean of the population of the $n$ squared distances between the two groups of $n$ points, minimized for all rotations, translations, and for all

correspondences $q$ allowed by the color of the points. When the number of colors is $k = 1$ (non-discernible particles model), it is the RMS spatial alignment without prefixed pairwise correspondence, and when $k = n$ (discernible particles model), it is the usual RMS spatial alignment, with prefixed (or implicit) pairwise correspondence. The RMS method is therefore extended to continuous and/or infinite sets of points, colored or not, provided that their inertia is finite.

## 4. Application of the colored mixture model to docking

It is proposed here to define a shape complementarity coefficient, or docking coefficient $\delta$, from a formal analog of equation (6), in which the variance of $(X_1 - X_2)' \cdot (X_1 - X_2)$ is used rather than its expectation. The intrinsic docking coefficient $\Delta_{12}$ is the minimzed docking coefficient, for all rotations $R$ and translations $t$ of the colored mixture $X_2$:

$$Z = (X_1 - X_2)' \cdot (X_1 - X_2), \tag{11}$$

$$\delta^2 = Inf_{\{W\}}\mathrm{Var}(Z), \tag{12}$$

$$\Delta_{12} = Min_{\{R,t\}}\delta. \tag{13}$$

Both colored mixtures $X_1$ and $X_2$ are assumed to have finite four-order moments, in order to ensure the existence of $\mathrm{Var}(Z)$.

We consider finite colored mixtures of finite discrete distributions, i.e. the conditions (a) and (b) in Section 2.3 are satisfied. Looking at equation (5) shows that minimizing $\mathrm{Var}(Z) = E[Z^2] - (E[Z])^2$, leads to minimizing a quadratic function under linear constraints, for which the probabilities $P(c)$ are fixed parameters, and the unknown quantities are either the probabilities $\tilde{W}_c(x_1, x_2)$, or the non-null elements of the rectangular bistochastic matrix $Q$ defined in section 2.3. As observed previously, the set of linear constraints is a closed bounded convex set, but, due to the presence of the quadratic term $(E[Z])^2$, we have a quadratic programming problem rather than a linear programming problem. Let $Q^*$ be a solution of this problem. As known, any point of the polytope of the constraints is expressible as a convex linear combination of the extremal points $Q_I$ of this polytope [10], thus $Q^* = \sum \lambda_I Q_I$, with $\lambda_I \geqslant 0$ and $\sum \lambda_I = 1$. The variance being obviously a concave function $f(Q)$ of the unknown matrix $Q$ (i.e. $-f(Q)$ is convex), it follows that $f(Q^*) \geqslant \sum \lambda_I f(Q_I)$, and since no quantity $f(Q_I)$ is lower than $\mathrm{Min}_{\{I\}} f(Q_I)$, it follows that $f(Q^*) \geqslant \mathrm{Min}_{\{I\}} f(Q_I)$, which is impossible unless the equality occurs, meaning that the minimized variance is reached at least on one of the extremal points of the convex polytope of the constraints.

We consider now the finite discrete colored mixture model, i.e. the five conditions (a)–(e) in Section 2.3 are satisfied. As for the similarity coefficient, computing the docking coefficient $\delta$ leads to enumeration of the permutation matrices $nQ$, $n$ being the number of points attached to the distribution of $X_1$

or $X_2$. Still denoting by $q$ the permutation associated to the permutation matrix $Q$, it follows that:

$$z_i = (x_{1_i} - x_{2_{q(i)}})' \cdot (x_{1_i} - x_{2_{q(i)}}) \tag{14}$$

$$\delta^2 = \mathrm{Min}_{\{q\}}\mathrm{Var}\{z_1; z_2; \ldots ; z_n\} \tag{15}$$

$$\Delta^2 = \mathrm{Min}_{\{R,t,q\}}\mathrm{Var}\{z_1; z_2; \ldots ; z_n\} \tag{16}$$

The proof of equation (15) or (16) is identical to that of equation (10) in Section 3, except that the mean of the population of the $n$ quantities $z_i$, is here replaced by its variance.

When all $n$ colors are different, there is a one-to-one pairwise correspondence between the two sets of $n$ points (discernible particles model). For this situation, using the variance was recently proposed [12] as a docking criterion. It is like the usual RMS spatial alignment method, except that the variance is used rather than the mean, and the docking is performed rather than the superposition. Except in [12], and despite its simplicity, our variance-based criterion was lacking in the literature (see [5] for a recent review).

Computing the intrinsic docking coefficient $\Delta$ in (16), is performed via enumeration of all correspondences (i.e. permutations) allowed by the colors of the points, after having computed the optimal translation and rotation. The analytical expressions of the optimal translation and the optimal planar rotation in (16) are known [12], and are extended to the general colored mixture model in appendices A and B.

Solving the full 3D docking problem requires both the optimal translation and the optimal spatial rotation. This is done for equation (16) with an iterative procedure based on a partly analytical solution, which leads to generating random initial rotations rather than random initial rotations plus translations [12]. Since the sampling of initial values is made for rotations only, the global minimum is retrieved at low computational cost, although usual docking procedures failed without some additional knowledge about the optimum to be computed, discarding the docking criterion [5].

The iterative numerical procedure solving the full 3D docking problem in [12] could be also extended to the general colored mixture model, provided that the joint distribution $W$ of the couple $(X_1, X_2)$ is fixed (see in appendix A.5 in ref. [6] how 3D rotations are handled). Then, the lower bound of the variance, taken over the set of the joint distributions $W$, has to be computed to obtain the intrinsic docking coefficient. As no specific algorithm is being actually devoted to this computation, standard numerical methods should be used.

## 5. Discussion and conclusion

The intrinsic docking coefficient $\Delta$ defined in equation (16) from the colored mixture model, may be normalized if needed. It has the major advantage of

being able to work both with discrete and continuous sets or distributions, even when one of the sets is discrete and the other is continuous. Infinite distributions are allowed, provided that their four-order moments are finite (e.g. gaussian mixtures).

A molecule, or a molecular fragment, may be modelized as a colored mixture of two distributions: the negative charges distribution and the positive charges distribution. It means that there are two colors, their prior probability being respectively proportional to the total negative charge and to the total positive charge of the fragment. For a neutral fragment, these two prior probabilities $P_1$ and $P_2$ are both equal to $1/2$. When non-neutral fragments are considered, the colored mixture model requires that the fragments receive a common ratio $P_1/P_2$.

Although the colored mixture model is adequate for similarity problems, it should be applied differently for docking. Since there is repulsion between charges of the same sign, the geometric docking has to be performed between distributions of opposite signs. Thus, docking two molecular fragments is performed such that the first color is attributed to the negative charges distribution in one of the fragments, and to the positive charges distribution in the other fragment, and conversely for the second color. A strictly geometric molecular model, i.e. without charges or with negative charges only, is of course handled with a unique color, and the docking is performed with two ordinary distributions, one modeling the static fragment, the other modeling the moving fragment.

Applying the colored mixture model to molecular docking needs a further assumption. Since molecular docking is often viewed as a local surface interaction, thus the geometric support of the underlying distributions have all a null volume.

The docking theory presented here has several advantages, including the simplicity of the docking creterion, but it also has a drawback: some situations lead to a null intrinsic docking coefficient, although they do not correspond to a satisfactory docking, from a practical point of view. An example is derived when a distribution is docked to its translated image. Obviously, $\Delta = 0$, but, except for a flat set (parallelism), the final docking may be such that the docked distributions are intersecting. It is proposed to add some further constraints in the minimization problem, in order to avoid this drawback. In other words, $\Delta = 0$ is a condition necessary to have a satisfactory docking, but sometimes it does not suffice.

The freeware DOG (*http://petitjeanmichel.free.fr/itoweb.petitjean.freeware.html*) performs the geometric docking between two sets of $n$ points pairwise associated [12]. Extending the program when the two sets have different cardinalities is under investigation, in order to have an automatic detection of the subsets having the best docking (relative to the size of the subsets). The optimal correspondence between the subsets has to be computed, as in the CSR similarity freeware [11] (same website as above).

## Appendix A

We look for the minimization of $\mathrm{Var}(Z)$ in equation (11), when $X_1$ is fixed and a translation $t$ is added to $X_2$, provided that the joint distribution of the couple $(X_1, X_2)$ is fixed:

$$Z = (X_1 - (X_2 + t))' \cdot (X_1 - (X_2 + t)), \tag{A.1}$$

$$V^* = \mathrm{Min}_{\{t\}}\mathrm{Var}(Z). \tag{A.2}$$

We set:

$$\tau = t + E[X2] - E[X1], \tag{A.3}$$

$$G = (X2 - E[X2]) - (X1 - E[X1]), \tag{A.4}$$

$$K_G = E[G \cdot G'], \tag{A.5}$$

$$\gamma = E[G \cdot G' \cdot G]. \tag{A.6}$$

Thus, $E[G] = 0$, $Z = (G + \tau)' \cdot (G + \tau)$, $K_G$ is the covariance matrix of $G$, and after some rearrangement:

$$\mathrm{Var}(Z) = \mathrm{Var}(G) + 4\tau' \cdot K_G \cdot \tau + 4\tau' \cdot \gamma. \tag{A.7}$$

Assuming that $K_G$ is invertible, the optimal translation $t^*$ is:

$$t^* = \tau^* - E[X2] + E[X1], \tag{A.8}$$

$$\tau^* = -K_G^{-1} \cdot \gamma/2, \tag{A.9}$$

$$V^* = \mathrm{Var}(G) - \gamma \cdot K_G^{-1} \cdot \gamma. \tag{A.10}$$

Equation (A.10) is formally identical to equation (9) in [12], which was obtained for the discrete colored model with all $n$ different colors.

When $K_G$ is not invertible, the difference of the centered colored mixtures $[X_2 - E[X_2]]$ and $[X_1 - E[X_1]]$ is subdimensional. The components of $\tau^*$ have to be found in a subspace, and its other components can take any value.

## Appendix B

We look for the minimization of $\mathrm{Var}(Z)$ in equation (11), when $X_1$ is fixed and $X_2$ is submitted to a rotation $R$, provided that the joint distribution of the couple $(X_1, X_2)$ is fixed:

$$Z = (X_1 - R \cdot X_2)' \cdot (X_1 - R \cdot X_2), \tag{B.1}$$

$$V^* = Min_{\{R\}}\mathrm{Var}(Z). \tag{B.2}$$

Moreover, we assume that the space is bidimensional, i.e. the planar rotation $R$, which is associated to the rotation angle $r$, is a linear combination of

the identity matrix $I$ (i.e. the null rotation), and of the matrix $\Pi$ associated to the +90 degrees rotation:

$$R = I \cdot \cos(r) + \Pi \cdot \sin(r), \tag{B.3}$$

$$Z = (X_1' \cdot X_1 + X_2' \cdot X_2) - 2(X_1' \cdot X_2) \cdot \cos(r) - 2(X_1' \cdot \Pi \cdot X_2) \cdot \sin(r). \tag{B.4}$$

Cov denoting the covariance operator, we set:

$$TT = \mathrm{Var}(X_1' \cdot X_1 + X_2' \cdot X_2), \tag{B.5}$$

$$CC = \mathrm{Var}(X_1' \cdot X_2), \tag{B.6}$$

$$SS = \mathrm{Var}(X_1' \cdot \Pi \cdot X_2), \tag{B.7}$$

$$CT = \mathrm{Cov}(X_1' \cdot X_2, X_1' \cdot X_1 + X_2' \cdot X_2), \tag{B.8}$$

$$ST = \mathrm{Cov}(X_1' \cdot \Pi \cdot X_2, X_1' \cdot X_1 + X_2' \cdot X_2), \tag{B.9}$$

and

$$CS = \mathrm{COV}(X_1' \cdot X_2, X_1' \cdot \Pi \cdot X_2). \tag{B.10}$$

It follows:

$$\mathrm{Var}(Z) = TT + 4 \cdot CC \cdot \cos^2(r) + 4 \cdot SS \cdot \sin^2(r) - 4 \cdot CT \cdot \cos(r) - 4 \cdot ST \cdot \sin(r) + 8 \cdot CS \cdot \sin(r)\cos(r) \tag{B.11}$$

$$\frac{1}{4}\frac{\partial \mathrm{Var}(Z)}{\partial r} = \cos(2r) \cdot (2 \cdot CS) - \sin(2r) \cdot (CC - SS) + \sin(r) \cdot CT - \cos(r) \cdot ST \tag{B.12}$$

Equation (B.12) is formally identical to equation (14) in [12], generalizing the result obtained for the discrete colored model with all $n$ different colors. The trigonometric expression in (B.12) is converted to a quartic polynomial of the unknown quantity $tg(r/2)$, and this polynomial is shown to have indeed real roots [12].

## References


[1] M.A. Johnson, J. Math. Chem. 3 (1989) 117–145.
[2] P.G. Mezey, J. Math. Chem. 7 (1991) 39–49.
[3] M. Petitjean, J. Comput. Chem. 16 (1995) 80–90.
[4] G.M. Maggiora, J.D. Petke and J. Mestres, J. Math. Chem. 31 (2002) 251–270.
[5] I. Halperin, B. Ma, H. Wolfson and R. Nussinov, Prot. Struct. Func. Gen. 47 (2002) 409.
[6] M. Petitjean, J. Math. Phys. 43 (2002) 4147–4157.
[7] B.S. Everitt and D.J. Hand, *Finite Mixture Distributions*(Chapman and Hall, London, 1981), ch. 1.
[8] S.T. Rachev, *Probability Metrics and the Stability of Stochastic Models*, (Wiley, New-York, 1991).

[9] S.T. Rachev and R.L. Rüschendorf, *Mass Transportation Problems*, Vol. I: *Theory* (Springer-Verlag, New-York, 1998).
[10] M. Minoux, *Programmation mathématique. Théorie et algorithmes*, (Bordas, CNET-ENST, Paris, 1983), Vol. 1 ch. 2, sect. 1.4 and 1.5.
[11] M. Petitjean, Comput. Chem. 22 (1998) 463–465.
[12] M. Petitjean, Internet Electron. J. Mol. Des. 1 (2002) 185–192, http://www.biochempress.com