Novel Treatment of Conformational Flexibility Using Interval Analysis

Zsuzsanna Szabó, Miklós Vargyas, and A. Peter Johnson*

ICAMS, School of Chemistry, University of Leeds, Leeds LS2 9JT, U.K.

Received August 15, 1999

A revolutionary new flexible conformational search technique is presented together with its application to geometrical docking. The algorithm is guaranteed to find all (even an infinite number of) solutions in a continuous manner. Examples are given using 2-(4′-amidinophenyl)pyruvate (APPA) docked to trypsin. Other potential applications are briefly discussed.

1. INTRODUCTION

The de novo ligand design program called SPROUT[1,2] has undergone continuous development at the University of Leeds for most of the past decade. Current research topics include methods for generation of ligands guaranteed to be easily synthesized and the rapid evaluation of members of combinatorial libraries for potential fit to a receptor.[3,4]

This paper focuses on some aspects of one of the most important steps in ligand generation, which is the docking of the growing chemical structure to the receptor's active sites.

Although the docking algorithm used in SPROUT[5] is one of the fastest current docking methods, it has a substantial limitation: it is a rigid docking procedure, with the caveat that several alternative rigid conformations are tried. In recent years there have been numerous attempts which aim to take into account the conformational freedom of the molecular structure being docked. The inherent difficulty with flexible docking is that the larger the conformational freedom of the molecular structure to be docked, the larger the search space to be explored, and hence the larger the execution time of the program. On the other hand, restricting conformational freedom may lead to loss of solutions, which is undesirable.

Our proposed approach to flexible docking allows full conformational freedom, which in our interpretation means that not only are certain discrete conformations of the molecule considered (e.g. those that have low strain energy) but *all possible conformations are evaluated on a continuous scale.*

1.1. Motivations. 1.1.1. Treatment of Rotatable Bonds. The approach to molecular flexibility in SPROUT (as in many other molecular modeling or design systems) represents a flexible molecule with several rigid conformations. In these structures the torsional angles of rotatable bonds are set to the theoretically or experimentally optimal values. For instance, unbranched alkyl torsions greatly prefer a trans orientation. A typical alkyl chain generated by SPROUT will not contain eclipsed interactions, because of the rules embedded in the system's knowledge bases.[7]

The main advantage of this approach is that it cuts down the size of the search space. The disadvantage is that the actual bound conformation of a ligand may have dihedral angles that differ significantly from these idealized values.[8,9] This means that (1) the natural ligand in its bound conformation cannot be generated in some cases (this is not a requirement, as SPROUT is a de novo design program, but the ability to generate the natural ligand in the conformation found in the crystal structure would provide a simple way to validate the program) and (2) it is unlikely, that the lowest energy conformation and the bound conformation of the ligand are identical or similar. In the case of phosphoethanolamine bound to phospholipase, two alkyl chains are found in the receptor's hydrophobic pocket (Figure 1). Most torsional angles for these chains lie between 160 and 180°, but one is 90° and one on the other chain is 100°.

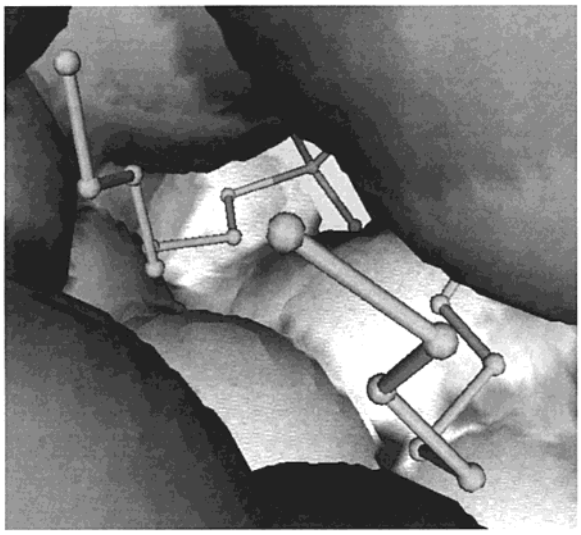


Figure 1. Phosphoethanolamine bound to phospholipase (PDB code 1POE). Both side chains have two planar parts, though one torsional angle in each is about 90° (pointing toward us).

Other approaches to conformational flexibility (like Monte Carlo methods[10] or genetic algorithms[11]) allow more exhaustive sampling of the search space, though—depending on the density of sampling—it is still more or less likely that some solutions are missed. Even if a docking method uses a continuous search technique, which does not restrict the search to some discrete representatives, for instance directed tweak,[12] the loss of solutions remains a problem, as only several local optima are found.

* To whom correspondence should be addressed. E-mail: pjohnson@chem.leeds.ac.uk.

10.1021/ci990105p CCC: $19.00
Published on Web 03/03/2000

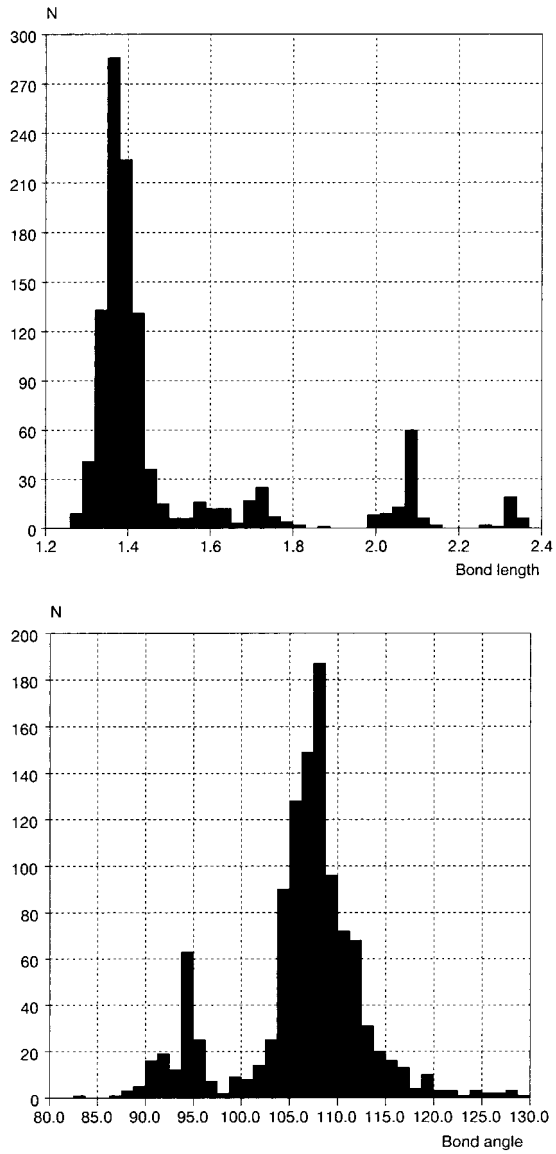


Figure 2. Five-membered aromatic ring geometries with unspecified atom type in CSD: (a, top) bond lengths; (b, bottom) bond angles.

1.1.2. Generic Structures. In SPROUT, besides specific chemical fragments, generalized fragments, so called templates, can also be used to build up molecular structures. In a generic template a generic atom can represent a number of alternative specific atom types; thus, a template represents a large set of structures. The advantage of using templates instead of real fragments is that it significantly reduces the size of the search space and helps to avoid a combinatorial explosion. However, the trouble with generalized templates is that typically their geometry corresponds to a carbon-only structure, and average carbon−carbon bond lengths, bond angles, and dihedral angles are used, for example 1.5 Å in the case of any single bond. In the case of five-membered aromatic rings, bond lengths are 1.4 Å, while bond angles are 108°. While these values represent carbon-only structures quite accurately, for heteroatom-containing structures, geometries are in some cases approximated poorly: as seen in the histograms, where no specific atom type is set, the variation in both bond length (Figure 2a) and bond angle (Figure 2b) is large. Not surprisingly, the largest deviations are found in sulfur-containing heteroaromatic compounds.

1.1.3. Hydrogen Bonding Sites. Hydrogen donor and acceptor sites of the receptor can be modeled in an easier fashion than other, more complex, global interactions (e.g. hydrophobic, π-stacking). Therefore, hydrogen bonding sites offer natural starting points for structure generation. In many systems hydrogen bonding sites are represented by cells of a discrete grid.[3] Starting fragments can be docked to any of these cells in some predefined discrete directions. In order to find out whether a fragment can be docked, all possible starting points and orientations have to be considered, which leads to a combinatorial explosion, as each starting fragment has many possible orientations and each one of these has to be extended with the next fragment.

In contrast, SPROUT represents hydrogen bonding sites as one continuous region: all points and all possible orientations that satisfy some given distance and angle conditions are taken into account too.[14] It means that a docked fragment has always only one starting orientation, which in later steps, as the structure grows, may be changed by the docking algorithm.

This continuous representation of hydrogen bonding sites was a very important milestone in the development of SPROUT: it cuts down the search space dramatically, but also gives rise to many more solutions than the original discrete sampling of orientations.

1.2. Aims. The success of the continuous hydrogen bonding site representation encouraged us to try to use the same idea, namely, a continuous description rather than discrete representatives, to tackle conformational flexibility: i.e. *using continuous dihedral angles and performing a full conformational search without sampling in order to find all possible solutions.*

In a similar fashion, uncertainties of generic as well as specific chemical structures could be modeled with continuous ranges of the most probable values.

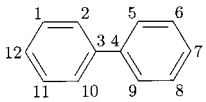
It has to be emphasized that targeting full, continuous conformational search in order to find all solutions and handling structural uncertainties is a very difficult problem: on the one hand the inherent complexity is very high, and on the other hand none of the commonly used methods in computational chemistry are appropriate to solve it (as those either use sampling or provide local solutions).

2. NOVEL APPROACH

A new technique was needed to tackle this problem: a search method that is capable of performing a continuous search on the whole conformational space without missing any possibility and a molecular structure representation that is suitable for this new algorithm.

2.1. Representing Flexible Molecules. In this system flexible molecules are described using local coordinates, i.e. bond lengths, bond angles, and dihedral angles, and all of these values are represented by one or more continuous ranges or, using mathematical terminology real intervals or a set of real intervals.

For instance in the case of biphenyl, the individual benzene rings are rigid; thus, only small deviations of bond lengths, bond angles, and dihedral angles are found in CSD ($\approx\pm0.03$ Å, $\approx\pm3°$, and $\approx\pm2.5°$, respectively (Figure 3)). This is why only tiny uncertainties are allowed in our representation of benzene. On the other hand, the central rotatable bond has



Bond lengths	
(1,12); (6,7); (7,8); (11,12)	[1.35Å, 1.41Å]
(1,2); (5,6); (8,8); (10,11)	[1.37Å, 1.41Å]
(2,3); (3,10); (4,5); (4,9)	[1.38Å, 1.42Å]
(3,4)	[1.48Å, 1.51Å]
Bond angles	
(1,12,11); (6,7,8)	[117.2°, 122.2°]
(5,6,7); (7,8,9); (10,11,12); (12,1,2)	[117.8°, 122.1°]
(1,2,3); (3,10,11); (4,5,6); (4,9,8)	[118.3°, 122.9°]
(2,3,10); (5,4,9)	[116.0°, 120.5°]
(2,3,4); (3,4,5); (3,4,9); (4,3,10)	[117.5°, 123.8°]
Torsional angles	
(5,6,7,8); (6,7,8,9); (10,11,12,1); (11,12,1,2)	[−2.02°, 2.42°]
(4,5,6,7); (7,8,9,4); (3,10,11,12); (12,1,2,3)	[−2.49°, 2.49°]
(1,2,3,10); (2,3,10,11); (8,9,4,5); (9,4,5,6)	[−0.51°, 0.68°]
(2,3,4,5)	[0.0°, 360.0°]

Figure 3. Interval representation of biphenyl. All data are derived from the Cambridge Database using 90% confidence intervals (i.e. 10% quantiles).

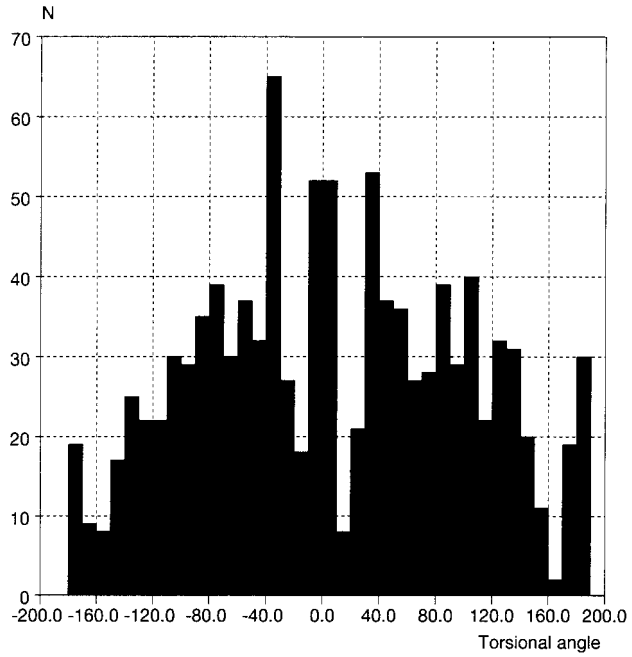


Figure 4. Distribution of torsional angle of the rotatable bond in biphenyl in CSD. It would be difficult and arbitrary to choose some representatives (e.g. 0°, ±30°).

almost complete torsional freedom: the median is 0.136°, and the mean is 6.659°, while the standard deviation of the sample is 93.003. The quantiles of a 90% confidence interval are at −118.122° and at 132.711° (Figure 4). This is why the dihedral angle of the rotatable bond joining the two benzene rings ranges from −180° to 180° in our representation (Figure 3).

The key point is that these intervals are not used just for mere storage purposes to describe the lower and upper limits for internal coordinates (as for instance the distance matrices used in distance geometry[15]) but rather are directly used in the calculations carried out by the docking algorithm.

2.2. Mathematics of Flexible Molecules. 2.2.1. Interval Analysis. The tool that looks eminently suitable for this work comes from a field of mathematics called interval analysis.[16−18] It is a fairly new field, which has been developing rapidly since the early 1960's.[19] It has found considerable application for so called "reliable computing".

Table 1. Values Calculated on an SGI O_2 When Evaluating Rump's Function Using Different Precision Real Arithmetics, Interval Arithmetics, and the Correct Value Calculated with Rational Arithmetic

real arithmetic	
64 bit	$-1.180\,591\,620\,71... \times 10^{21}$
128 bit	1.172 603 940 05...
interval arithmetic	
64 bit	$[-7.083\,549\,724\,304... \times 10^{21}, 4.722\,366\,482\,86... \times 10^{21}]$
128 bit	$[-1.310\,708\,273\,96... \times 10^{5}, 1.172\,603\,940\,05...]$
correct value	$-0.827\,396\,059\,946\,821\,3 \pm 10^{-16}$

Computer calculations with real numbers always lead to computational errors because the result of any calculation is rounded to the nearest representable value. For example when the following function of two variables (example of Rump[20])

$$f(x, y) = 333.75y^6 + x^2(11x^2y^2 - y^6 - 121y^4 - 2) + 5.5y^8 + \frac{x}{2y}$$

is evaluated in $x = 77\,617$, $y = 33\,096$ on a computer, the result varies tremendously, depending on the precision of the real arithmetic used, and none of these results are correct (see Table 1). Using interval analysis ensures that the correct value lies in the interval that we get as a result, as all the calculations are rounded in both directions. In the example, the result interval is very wide even in the extended precision case, which indicates that the rounding error of calculations is very significant. The correct value is contained in the result intervals.

Nowadays interval analysis is mainly used as a modeling framework for representing uncertainty, e.g. in systems where imprecise measurements are involved. There are several applications in geology,[21] engineering,[22] robotics,[23] and computer graphics.[24,25]

Interval analysis is an extension of mathematical analysis of real numbers. For example addition and multiplication are defined the following way:

$$[a, b] + [c, d] = [a + b, c + d]$$

$$[a, b]\cdot[c, d] = [\min(ac, bc, ad, bd), \max(ac, bc, ad, bd)]$$

Example:

$$[1, 5] + [-3, 7] = [-2, 12]$$

$$[1, 5]\cdot[-3, 7] = [-15, 35]$$

The result of an operation on intervals is an interval which contains each value that can be obtained by applying the same operation on any real value taken from the operand intervals; this is the *enclosing principle.* In mathematical terms

$$F:\ \mathrm{IR}^n \to \mathrm{IR}^m$$

is an *interval extension* of $f:\ \mathrm{R}^n \to \mathrm{R}^m$ iff

$$\forall X \in \mathrm{IR}^n \quad \text{and} \quad \forall x \in X:\ f(x) \in F(X)$$

where R^k is the space of k-dimensional real vectors and IR^k is the space of k-dimensional interval vectors. Looking at

the examples, if any value from −3 to +7 is added to any value from +1 to +5, than we always get a result between −2 and +12. If we multiply them, the result is always between −15 and +35.

In these simple examples the result intervals are tight: they contain only those values that can be obtained by applying the corresponding real operations on every value pair in the interval operands. It is not necessarily so with other functions, where the result interval may contain other values. In general the only requirement for an interval function is the enclosing principle mentioned above.

2.2.2. Advantages and Disadvantages. Using interval analysis has two important advantages. It is guaranteed to find all solutions to a problem, even an infinite number of solutions, if they are on a bounded set. Moreover, it is guaranteed to find the global optimum of an optimization problem.[18,26] This cannot be achieved by any of the commonly used point based methods.

These advantages have their cost: interval methods are more time and space consuming than point based methods. On the other hand this extra cost might be quite justifiable where noninterval methods fail to solve the problem.

In most circumstances there is a problem inherent in the use of interval analysis, which can reduce its effectiveness significantly. If a variable is repeated several times in an expression, then evaluating this expression may give a resulting interval which is too wide because all the occurrences of the variable are treated as independent variables. This is the so called *dependency problem.* We take a simple example:

$$f(x) = x(x - 1)$$

Variable x occurs twice in f, and if f is evaluated on $x = [-2, 3]$

$$f([-2, 3]) = [-2, 3]\cdot([-2, 3] - 1) = [-2, 3]\cdot[-3, 2] = [-9, 6]$$

the result is [−9, 6]. However, the tightest result would be [−0.25, 6]. If f is rearranged, $f(x) = x(x - 1) = x^2 - x$, then we get

$$f([-2, 3]) = [-2, 3]^2 - [-2, 3] = [0, 9] - [-2, 3] = [-3, 11]$$

The result is different, but still very wide.

2.2.3. Tackling the Dependency Problem. There are several approaches to tackle the dependency problem. In this work we used a specific interval method called *affine arithmetic*,[25] where every interval x is associated with an affine expression of the following form:

$$x = x_0 + \sum_{i=1}^{n} x_i \epsilon_i$$

where x_0 is the center of interval x, x_i is a real number, and ϵ_i lies in the interval [−1, 1] (for all $i = 1, ..., n$). Using this form, the dependency between the variables of an expression is not lost, and thus more accurate results are obtained.

Other methods which tackle the dependency problem include generalized interval arithmetic[27] and geometrical objects called zonotopes.[28]

2.3. Model of Flexible Docking. Mathematical models consist of three main categories: parameters, variables, and constraints. It is important to note that different models are constructed depending on whether a feature of the modeled phenomenon is described by variables or constraints. For example, in our problem, the hydrogen bonding situation can be described as constraints applied to the position of the built structure's bonding atoms or by introducing variables to represent the constrained angles and bond lengths and using the allowed ranges as the initial values of these variables. In our work we have created several models, and our current one describes bond lengths, bond angles, and the dihedral angles of nonrotatable bonds as parameters. The hydrogen bonding situation is represented by six variables: three dihedral angles, two bond angles, and the length of the hydrogen bond. This results in the same representation as in SPROUT.[14] The dihedral angles of the rotatable bonds and the coordinates of the ligand's atoms are also variables. Boundary fit and allowed minimal van der Waals distances are the constraints applied. These constraints are the only ones which have been introduced because our aim is to perform docking in two stages: (1) geometrical docking to find all feasible docked conformations; (2) energy minimization to find good binding. Geometrical docking is modeled with a system of interval equations:

$$F(X) = B, \quad F: \mathrm{IR}^n \rightarrow \mathrm{IR}^m, \quad X \in \mathrm{IR}^n, \quad B \in \mathrm{IR}^m$$

It is intended that energy minimization will be modeled with interval optimization.

In order to describe the system of equations of the first stage, we denote $p_i = (x_i, y_i, z_i)$ as the atom coordinates of the ith atom of the built structure and $d(p_i, p_j)$ the Euclidean distance between two atoms. Then the system consists of the following four classes of equations: (1) the distance between two atoms i, j modeling three different constraints (bond length of bonded atoms, minimal van der Waals distance of nonbonded atoms, and the boundary fit (minimal distance between the atoms of the structure and the receptor))

$$d^2(p_i, p_j) - d_{ij}^2 = 0$$

where d_{ij} is the required distance; (2) the distance between two atoms i, k that have a common neighbor atom j modeling the bond angle constraint between i, j, k

$$d^2(p_i, p_k) - (d_{ij}^2 + d_{jk}^2 - 2d_{ij}d_{jk}\cos\alpha) = 0$$

where d_{ij} and d_{jk} are the bond lengths between i, j and between j, k and α is the bond angle between i, j, k; (3) the distance between two atoms i, l that have bonded neighbors j, k modeling the dihedral angle constraint between i, j, k, l

$$d^2(p_i, p_l) - ((d_{jk} - d_{ij}\cos\alpha - d_{kl}\cos\beta)^2 + (d_{kl}\sin\beta - d_{ij}\sin\alpha\cos\phi)^2 + d_{ij}^2\sin^2\alpha\sin^2\phi) = 0$$

where d_{ij}, d_{jk}, and d_{kl} are the bond lengths between i, j, between j, k, and between k, l, α is the bond angle between i, j, k, β is the bond angle between j, k, l, and ϕ is the dihedral angle between i, j, k, l; (4) atom coordinates of atom l expressed from the internal coordinates and from the coordinates of the other three atoms of the structure (i, j, k),

Chart 1

$$\mathbf{T} = \begin{pmatrix} x_j - x_k & \dfrac{d_{jk}(x_i - x_k) - (x_j - x_k)(d_{jk} - d_{ij}\cos\alpha)}{d_{ij}d_{jk}\sin\alpha} & \dfrac{A}{[A^2 + B^2 + C^2]^{1/2}} \\ y_j - y_k & \dfrac{d_{jk}(y_i - y_k) - (y_j - y_k)(d_{jk} - d_{ij}\cos\alpha)}{d_{ij}d_{jk}\sin\alpha} & \dfrac{B}{[A^2 + B^2 + C^2]^{1/2}} \\ z_j - z_k & \dfrac{d_{jk}(z_i - z_k) - (z_j - z_k)(d_{jk} - d_{ij}\cos\alpha)}{d_{ij}d_{jk}\sin\alpha} & \dfrac{C}{[A^2 + B^2 + C^2]^{1/2}} \end{pmatrix}$$

```
put initial search region into L
while L ≠ 0
   remove first box X from L
   if X is proven not to contain solution  then
      reject X
   else if X is proven to contain only solutions  then
      accept X
   else if X can be reduced  then
      reduce X
      put back X into L
   else
      subdivide X into X1, X2, ..., Xn
      put X1, X2, ..., Xn into L
   endif
endwhile
```

Figure 5. Abstract branch-and-bound algorithm.

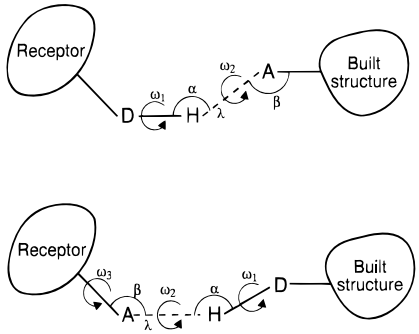


Figure 6. Representing the hydrogen bonding situation: (a, top) hydrogen donor site; (b, bottom) hydrogen acceptor site. ω_i denotes full rotational freedom ([0°, 360°]) around the corresponding axes. Note that ω_3 is also [0°, 360°], even if the corresponding bond is a double bond; this is necessary to obtain all allowed positions of the donor H atom. The default values for other parameters are as follows: α = [135°, 180°], β = [90°, 180°], and λ = [1.6 Å, 2.2 Å].

where *i* and *j*, *j* and *k*, *k* and *l* are bonded (this constraint class was introduced to enhance the effectiveness of the applied interval method)

$$p_l - \left(p_k + \mathbf{T} \begin{pmatrix} c\cos\beta \\ c\sin\beta\cos\phi \\ c\sin\beta\sin\phi \end{pmatrix} \right) = 0$$

where **T** is the transformation matrix in Chart 1.
In this formula (*A*, *B*, *C*) is the normal vector of the plane that lies on p_i, p_j, and p_k and can be calculated as follows:

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} = (p_i - p_j) \times (p_k - p_j)$$

d_{ij}, d_{jk}, d_{kl}, α, β and ϕ are defined as in class 3 above.

2.4. Algorithm. The algorithm that is used during the conformational search is a *branch-and-bound* algorithm[29] and is shown in Figure 5 in a very generic form. Here *L* is a container (e.g. a list) of all of those variables' values that may lead to solutions. In our case, these are multidimensional interval vectors or, using a simpler term common in interval analysis, multidimensional boxes[30] (or hyperboxes, hypercubes). In the beginning the whole search region (one or several initial boxes of variable values) is put into *L*, since it is supposed to contain solutions.

During execution, a box can be discarded, accepted as a solution, reduced, or subdivided. It is discarded if it is proven not to contain solutions, i.e. if the constraints' values, evaluated in the given box, do not contain any of the required values (in the case of a system of equations, $F(X) \cap B = \emptyset$, where *X* is the box). The box is accepted as a solution, if it contains only solutions (within a given error); i.e. the constraint's values evaluated in the box contain only required values ($F(X) \in B$). It may be reduced by removing proven nonsolutions. There are several techniques in interval analysis to reduce a box. In most of the cases the system has to be linearized, which is done by either calculating the functions' derivatives[17,18] or giving interval linear approximations of the function.[25] The reduction step of the algorithm is always the most time consuming. If the reduction attempt fails, then the box is subdivided into subboxes, and the subboxes are all stored in *L* for further processing. There are also several approaches to decide which dimension of the box should be subdivided; the most simple one is to divide the one with the biggest diameter.

The algorithm terminates when all boxes contained in *L* are accepted or discarded; thus, *L* remains empty. It is proven that the algorithm will always terminate.[18]

It has to be noted that in our work every effort has to be made to reduce a box, instead of subdividing it, as subdivisions lead to exponential growth of the size of *L*. In order to save time, several reduction methods are applied, starting with the fastest one and following with the more time consuming but more efficient ones. The fast methods include steps of the interval Gauss−Seidel iteration; the slowest is a solution of two linear programming problems for each variable to find the tightest possible bounds.

3. RESULTS AND DISCUSSION

To illustrate the method, we have docked APPA to five hydrogen bonding sites of Trypsin,[31] allowing all chemically reasonable situations at the sites and full conformational freedom in APPA's three rotatable bonds. To validate the algorithm, we need to answer the question, can our docking method find the bound conformation of APPA as found in the crystal structure, among all other solutions? The bound conformation of APPA is considered to have been found by the algorithm, if among all solutions there is at least one,

where the atom coordinate boxes are reasonably small and still enclose each atom of the bound conformation as found in the crystal structure.

The variables of the system are defined by the five target sites and the flexibility of the ligand. They are listed below together with their initial values.

(1) flexibility at the two H-donor sites, each six variables (Figure 6a)

3 rotatable bonds: [0°, 360°]

2 bond angles: [90°, 180°], [135°, 180°]

length of the hydrogen bond: [1.6 Å, 2.2 Å]

(2) flexibility at the three H-acceptor sites, each five variables (Figure 6b)

2 rotatable bonds: [0°, 360°]

2 bond angles: [90°, 180°], [135°, 180°]

length of the hydrogen bond: [1.6 Å, 2.2 Å]

(3) three rotatable bonds of the ligand: [0°, 360°]

(4) coordinates of the ligand atoms: the coordinates of a large box ([−8.6393, 8.7512], [−6.2439, 4.9591], [−5.6456, 4.5538]) (the dimension of the diameter of the intervals is measured in angstroms), that contains the binding pocket, as originally the ligand atoms can be anywhere inside the cavity

Using the constraint categories defined in 2.3, the system consists of the following.

(1) altogether 1451 distance constraints, of which

22 bond length constraints

74 van der Waals distance constrains

1355 intermolecular distance constrains
(boundary fit, 89 receptor atoms are involved)

(2) 32 bond angle constraints

(3) 44 dihedral angle constraints

(4) 264 expressed atom coordinate constraints

During processing these values are reduced and subdivided to satisfy all constraints until all possible variable values are considered, all acceptable solutions are found, or all values are discarded, and thus it is proven that there are no solutions.

Without subdivisions, atom coordinate boxes can be reduced significantly, but they are still too large. After about 50 subdivisions, atom coordinates are reasonably small, and the conformation shown in Figure 7a is one that contains the bound conformation of APPA. This validates the algorithm. However, it has to be noted that 50 subdivisions results in an immense search tree.

When the flexibility is reduced, allowing only one-sixth of the full dihedral angle intervals, half of the bond angle intervals, and two-thirds of the bond length tolerance (at the hydrogen bonding sites), atom coordinate boxes can be reduced much faster. Figure 7b is the conformation reached using only reduction steps, and Figure 7c shows a sample conformation after 10 subdivisions; of course this is not yet the final solution.

In order to obtain lucid illustrations of the final results of a full geometrical docking, where all the solutions within 0.5 Å accuracy are found, the size of bond angle and dihedral angle intervals was reduced further by half. The results are shown in Figure 8.

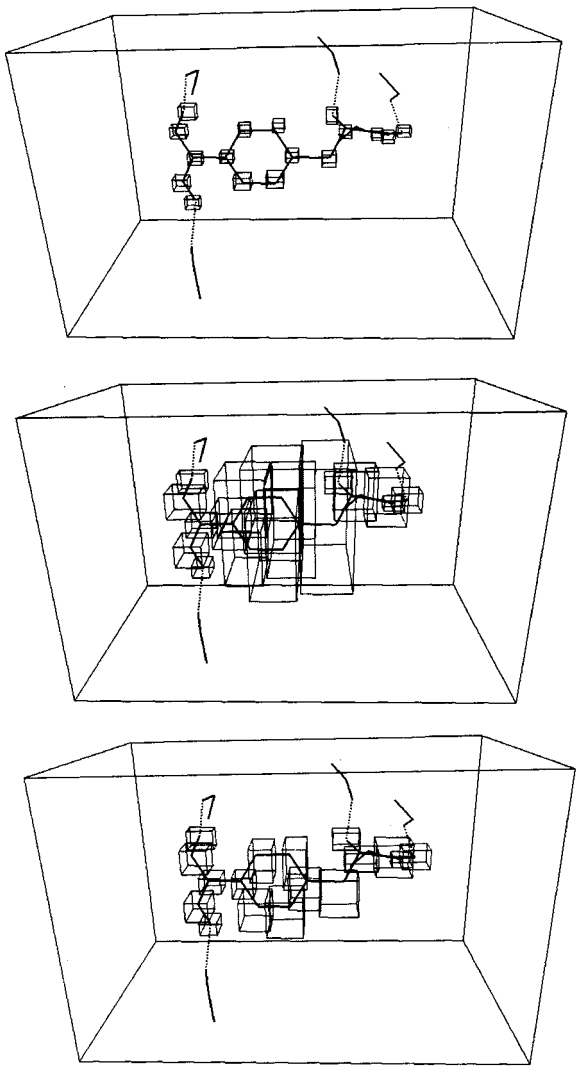


Figure 7. Docking APPA to five hydrogen bonding sites of trypsin (two donors, O_γ Ser 190, O Gly 219, and three acceptors, N Ser 195, N Gly 193, N_{ϵ_2} His 57). The drawn molecule is APPA as found in the crystal structure. It is used purely for display purposes and is not used by the algorithm. The large box contains the binding pocket (about 90 heavy atoms of the receptor) and acts as the initial position of the atoms of APPA to be docked: (a, top) a possible conformation found after 50 subdivisions when allowing full flexibility (as part of all solutions, this partial solution contains the conformation of APPA which was determined in the crystal structure; the small boxes enclose all possible spatial positions of the corresponding atom centers; (b, center) all allowed conformations, reached without subdivisions when the flexibility is reduced (as this is not the final solution (i.e. still poor approximation), nonsolutions are also contained in the boxes); (c, bottom) sample conformation after 10 subdivisions when the flexibility is reduced (this gives much better approximation of the whole solution space).

The main advantage of our method is that it is capable of finding all possible solutions by carrying out a full, continuous conformational search. As a consequence, if the program terminates with no solutions, then it is guaranteed that there are no solutions to the problem, since all possibilities were considered. *No other known docking method exhibits this feature.*

Other unique features include the following: (1) those constraints that have already been satisfied do not need to be rechecked;[32] (2) uncertainties of both the ligand and the

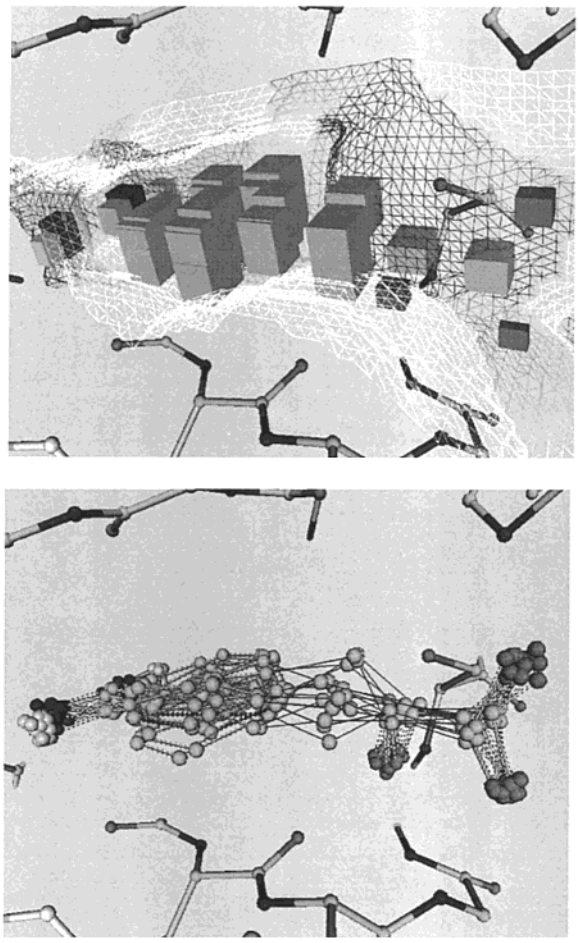


Figure 8. All solutions within 0.5 Å error in atom center position (using reduced flexibility). Solution boxes are displayed in the receptor's cavity: (a, top) boxes represent all possible positions of APPA's atom centers; (b, bottom) these are the same results in more common representation for the sake of easier visualization (these sample conformations are arbitrarily taken by sampling the entire solution space).

protein structure can be taken into account by using appropriate intervals for bond lengths and bond angles.

Most importantly, this approach seems to be generic enough to be applied to other problems, some of which are discussed below.

The only obvious disadvantage is the long running time, and work is proceeding with the goal of reducing it to an acceptable level. The current version runs in parallel on a network of several Silicon Graphics workstations. A central R5000 O_2 runs the main branch-and-bound algorithm, evaluates the functions, stores the search space, and performs reduction steps and subdivisions. Since the most exhaustive reduction step requires the solution of many optimization subtasks, these run in parallel. In our experiments, we used 10 SGs (with various processors) that run an interior point method linear programming (LP) problem solver.[33]

One such reduction step requires 160−180 LP problems to be solved (depending on the number of hydrogen sites and rotatable bonds). One LP otimization takes half a second on average; thus, a full reduction takes about 45 s. The total running time of the program depends heavily on the size and the nature of the problem. More constrained problems, where the solution space is small (especially in the extreme case, where there is no solution at all), are solved relatively fast (several minutes on the configuration mentioned above). However, a less constrained problem, where the solution space is large, needs several hours.

Understanding the importance of strict constraints led to the recognition of the deficiency of purely geometrical docking: in most cases, the conformations of the ligand that fit the receptor's cavity covers a fairly large volume. Finding this space with good accuracy (e.g. 0.5 Å) requires long running times. Thus, it is important to consider secondary constraints (e.g. binding energy) to reduce the conformational space.

Due to the immense complexity of this problem, some simplifications have been applied in order to find out if the primary aim (i.e. full continuous conformational search) can be reached: (1) uncertainties in bond lengths and bond angles have been introduced as parameters, not as variables; (2) alternative mappings of ligand atoms to receptor sites are not determined by the algorithm, though interval analysis based docking potentially offers some smart ways to eliminate the combinatorial nature of the mapping problem, and this is currently under investigation; (3) we have been working only on the first phase of docking, i.e. only geometrical docking is performed, and no account is taken of binding energy. Virtually any existing method of estimating binding energy (including MM, empirical scoring functions) can be combined with this method, although all of these energy calculations would have to be extended to deal with intervals.

4. CONCLUSION

A novel algorithm to calculate all those 3D conformations of a flexible molecule that satisfies a set of given constraints is presented. The latest results prove the feasibility of the ideas presented, although whether this new kind of docking has practical applicability is still an open question. Beside massive parallelization, ways to overcoming the problem of the long running time include alternative methods for improving the accuracy of affine approximations and extending the set of available reduction methods with faster algorithms.[34]

However, it is very likely that there are many applications where this approach could be clearly superior to other existing methods, e.g. in energy minimization, where finding the global energy minima would be guaranteed, and all conformations within a predefined window of energy could be found as well.

Other problems, where the inherent uncertainties play a crucial role, like structure elucidation by NMR or X-ray and pharmacophore identification, are probably also good candidates for further research.

ACKNOWLEDGMENT

We are grateful to our colleague Krisztina Boda, who helped us with searches of the Cambridge Database. We had hours of fruitful discussions with Zsolt Zsoldos, who is owed many thanks for his inspiring work on SPROUT. Two of the authors were supported by ORS and Tetley&Lupton Awards.

REFERENCES AND NOTES

(1) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A program for structure generation. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 127−153.

(2) Gillet, V.; Newell, W.; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: Recent Development in the de Novo Design of Molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207−217.
(3) Boda, K. De Novo Design of Synthetically Accessible Ligands. University of Leeds, Ph.D. work, unpublished.
(4) Marchaland, J. F. VLSPROUT: A Structure-Based Virtual Library Screening System. To be published.
(5) Zsoldos, Zs. New Method for *de Novo* 3D Structure Design. Ph.D. Thesis, University of Leeds, 1996; Chapter 4.
(6) The Cambridge Structural Database (CSD) was searched by Quest version 2.3.7; histograms were produced by X-Vista v2.1. CSD and associated search software are available at the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk/prods/csd.html).
(7) By default 60, 120, and 180° are generated, but the user can specify arbitrary torsional angles.
(8) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational Changes of Small Molecules Binding to Proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411−428.
(9) Jorgensen, W. L. Rusting the Lock and Key Model for Protein-Ligand Binding. *Science* **1991**, *254*, 954−955.
(10) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrated to Proteins by Simulated Annealing. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195−202.
(11) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and Validation of a Genetic Algorithm for Flexible Docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(12) Hurst, T. Flexible 3D Searching: The Directed Tweak Technique. *J. Chem. Inf. Sci.* **1994**, *34*, 190−196.
(13) Böhm, H. J. The computer program LUDI: A new method for the de novo design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61−78.
(14) Zsoldos, Zs. New Method for *de Novo* 3D Structure Design. Ph.D. Thesis, University of Leeds, 1996; Chapter 2.
(15) Crippen, G. M. Rapid Calculation of Coordinates from Distance Matrices. *J. Comput. Phys.* **1978**, *26*, 449−462.
(16) Alefeld, G.; Herzberger, J. *Introduction to Interval Computations*; Academic Press: Orlando, FL, 1983.
(17) Hansen, E. R. *Global Optimization Using Interval Analysis*; Dekker: New York, 1992.
(18) Neumaier, A. *Interval Methods for Systems of Equations*; Cambridge University Press: Cambridge, U.K., 1990.
(19) Moore, R. E. *Interval Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1965.
(20) Rump, S. M. Algorithm for Verified Inclusions: Theory and Practice. In *Reliability in Computing*; Moore, R. E., Ed.; Academic Press: Boston, MA, 1988; pp 109−126.
(21) Doser, D. I.; Crain, K. D.; Baker, M. R.; Kreinovich, V.; Gerstenberger M. C. Estimating Uncertainties for Geophysical Tomography. *Reliab. Comput.* **1998**, *4*, 241−268.
(22) Birdie, T. R.; Surana K. S. The Use of Interval Analysis in Hydrologic Systems. *Reliab. Comput.* **1998**, *4*, 269−281.
(23) Morales, D.; Son, T. C. Interval Methods. in Robot Navigation. *Reliab. Comput.* **1998**, *4*, 55−61.
(24) Brito, A. E.; Kosheleva, O. Interval + Image = Wavelet: For Image Processing under Interval Uncertainty, Wavelets Are Optimal. *Reliab. Comput.* **1998**, *4*, 283−290.
(25) de Figueiredo, L. H.; Stolfi, J. Adaptive Enumeration of Implicit Surfaces with Affine Arithmetic. *Comput. Graphics Forum* **1996**, *15*, 287−296.
(26) Hansen, E. R. An Overview of Global Optimization Using Interval Analysis. In *Reliability in Computing*; Moore, R. E., Ed.; Academic Press: Boston, MA, 1988; pp 289−307.
(27) Hansen, E. R. A Generalized Interval Arithmetic. In *Interval Mathematics*; Nickel, K. L., Ed.; Springer-Verlag: New York, 1975; pp 7−18.
(28) Kühn, W. Rigorously Computed Orbits of Dynamical Systems without the Wrapping Effect. *Computing.* **1998**, *61*, 47−67.
(29) Kearfott, R. B. A Review of Techniques in the Verified Solution of Constrained Global Optimization. In *Applications of Interval Computations*; Kearfott, R. B., Kreinovich, W., Eds.; Kluwer: Dordrecht, The Netherlands, 1996; pp 23−60.
(30) These boxes are elements of an abstract vector space. However, if for instance three variables (out of many others) are assigned to the three Cartesian coordinates of an atom in the *x, y, z* space, then these three variables, as a subbox of a larger hyperbox, form a "real" three-dimensional box in the Euclidean space. On the other hand, if three independent variables represent the torsional angles of three different rotatable bonds, then these do not represent a "real" three-dimensional box, although a corresponding box in the *x, y, z* space could be assigned, but this box does not "exist" in strict terms.
(31) This method involves user selection of a subset of all the hydrogen bonding sites of the protein and mapping of ligand atoms to these sites. Therefore it could be argued that the system described here is not an unbiased docking procedure. However further work designed to overcome this limitation has been completed and the resulting system is being tested.
(32) This feature has importance in structure generation, making the growing process more efficient.
(33) Mészáros, C. *The BPMPD interior point solver for convex quadratic problems*; WP 98-8; Computer and Automation Research Institute, Hungarian Academy of Sciences: Budapest, 1998.
(34) These faster reduction steps are usually less effective; i.e. there are cases where they would fail to perform reduction when slower, more comprehensive methods would succeed. However, in many cases even simple reduction procedures may be adequate, thus avoiding the need for the slower, exhaustive methods.

CI990105P