Flexibility, Conformation Spaces, and Bioactivity

Oren M. Becker,* Yaakov Levy, and Orr Ravitz

Department of Chemical Physics, School of Chemistry, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel

Received: July 1, 1999

Conformation constraints and molecular flexibility strongly affect the bioactivity of flexible molecules. The present study offers a new conceptual framework, as well as a practical quantitative procedure, for discussing and quantifying these effects. The theory is formulated in terms of weighted overlaps between the volume in conformation space occupied by the flexible ligand and the pre-prescribed conformational requirements imposed by the host molecule ("region of bioactivity"). From this theory a quantitative structure activity relationship (QSAR) type descriptor, which quantifies the effect of conformation constraints on bioactivity, was derived and the resulting model was shown to be in excellent correlation with the observed activity of the molecules. Three characteristic scenarios for the relationship between flexibility and bioactivity are outlined and demonstrated in realistic systems: conformationally constrained alanine hexapeptides, a series of substance P analogues, and a set of conformationally constrained Arg−Gly−Asp containing peptides.

Introduction

Conformation constraints and molecular flexibility are known to have a very strong effect on the activity (in particular binding affinity) of flexible molecules. The binding affinity of peptides, as well as that of other flexible compounds, is often altered by conformation constraints such as cyclization, enantiomeric substitutions, and the introduction of stereochemically constraining chemical groups. Based on this observation, medicinal chemistry optimization of lead compounds often proceeds along two avenues. The first avenue focuses on chemical modifications (changing the chemical properties of the molecule), while the second proceeds through the application of conformation constraints. There are many examples for conformationally constrained analogues that are more bioactive or more specific than the original lead molecules. For example, application of various conformation constraints to Arg−Gly−Asp containing peptides, which are the primary recognition site for cell adhesion, affects both their binding affinity and specificity.[1] Another example is the dramatic effect of enantiomeric substitution on the binding of substance P analogues to NK1 receptor.[2,3] Such conformational considerations also play a major role in the development of peptido-mimetic drugs.[4] Similar considerations of flexibility and conformation constraints are, of course, also applicable to many chemical design problems in nonpharmaceutical applications. There are also cases in which the structure of both host and ligand may change upon binding. A certain amount of flexibility is required in these cases for binding, rendering too rigid ligands less effective.

The recognized need to account for molecular flexibility in drug development has been a strong motivation for recent developments in computer aided drug discovery, both in the field of quantitative structure activity relationship (QSAR) and in the field of structure-based molecular docking. Recognizing the role of flexibility during the docking process (both in the ligand and in the receptor) led to a recent surge in activity that resulted in many new "flexible docking" methodologies.[5] Most of these methods address flexibility either by representing the docked molecule as a set of molecular replicas, reflecting different possible conformations, or by a stepwise "anchor and grow" approach, in which molecular fragments are optimally linked together inside the binding site. Molecular flexibility is even harder to account for in the context of QSAR, much because in this case the structure of the binding site is often unknown. Classical QSAR is based on correlating the chemical properties of the molecule (e.g., charge and lipophilicity) with activity, using a large number of chemical "molecular descriptors". In recent years 3D-QSAR methods, which take into account structural similarity, have become standard tools. Nonetheless, these methods rely on knowing the "structure" for each molecule in the data set. For flexible molecule such "structures" are not well defined. Current QSAR methodology, in general, cannot account for the important properties of "flexibility" and "conformational entropy", which clearly play a role in determining the binding affinity. Only when the structure of the binding site is known can conformational entropy be introduced into QSAR through free energy calculations.[6] The lack of quantitative QSAR "descriptors" for the overall effect of conformation constraints (not just stereochemical changes of the lowest energy conformation), especially in cases where the structure of the receptor is unknown, clearly limits the scope of QSAR. Recently Hopfinger et al.[7] introduced 4D-QSAR, in which flexibility is accounted for by assigning population probabilities to the Cartesian 3D grid already used in 3D-QSAR.

In the present study we outline a conceptual framework for discussing the role of molecular flexibility and conformation constraints in bioactivity. Three representative scenarios explain the relation between flexibility, constraints, bioactivity, and specificity in terms of "molecular conformation spaces" (rather than the molecules' specific 3D structures). A computational procedure which quantifies these concepts is applied to three sets of conformationally constrained peptide families (alanine hexapeptide analogues, substance P analogues, and RGD-containing septapeptides) and links them to the proposed

* To whom correspondence should be addressed. E-mail: becker@sapphire.tau.ac.il.

10.1021/jp992268m CCC: $19.00
Published on Web 02/09/2000

scenarios. Finally, the results of these quantitative analyses, which attest to the validity of the suggested concepts, are discussed.

Conceptual Framework

Many computer aided drug design approaches, both of the docking type and of the 3D-QSAR type, try to account for molecular flexibility by replacing the single structure used in the standard “nonflexible” applications with a relatively small set of molecular replicas, each with a different conformation. These conformationally distinct replicas are generated either consistently (e.g., combinatorically varying all rotatable dihedral angles) or by using sampling procedures. Since the computation treats each replica as an independent molecule, the number of “molecules” to analyze often increases by a factor of 10 or more. This increases the computational load and reduces the usefulness of these tools for high throughput screening. Moreover, replacing a single conformation by a small set of alternative conformations does not reveal much about the molecular properties of “flexibility” and conformational entropy. Such conformation samples also do not address the many important questions which relate to the whole “world of conformations” available to the molecule; e.g., How do conformational constraints affect molecular flexibility? To what degree do individual constraints reduce or increase flexibility? What degree of conformational flexibility is required for a molecule to bind efficiently? What is the relative flexibility of analogous molecules and how does that affect their bioactivity?

In principle, questions about molecular flexibility should be discussed in terms of the size and shape of the corresponding molecular “conformation space”. By definition, flexible molecules adopt more conformations than their nonflexible counterparts. Namely, the volume they occupy in conformation space is larger. Rigid molecules, on the other hand, are restricted to small volumes in conformation space since only a small number of conformations are available to them.

The concept of “conformation space volume” can be used as a framework for discussing the relative binding affinity of flexible molecules (the term binding affinity is used here in the same, somewhat loose, manner it is used in experimental bioactivity studies). The *conformational aspect* of the binding affinity (docking) is an interplay between the predefined set of conformations that can, in principle, fit the binding site and the actual set of conformations that the molecule can adopt. For rigid molecules the question of conformational compatibility reduces to a yes/no answer. Either the rigid molecule fits the binding site (to within some fitting criterion) or it does not. For flexible molecules the question becomes statistical and depends on the percentage of conformations that can fit into the binding site. The larger this percentage the higher will the binding affinity be. More accurately, a Boltzmann weighted percentage should be used to reflect the likelihood for the molecule to be in any of its possible conformations.

Let us define the “occupied volume in conformation space”, V_{conf}, as the whole set of conformations that can be adopted by the ligand under physiological conditions. Let us further define the “region of bioactivity”, R_{bio}, prescribed by the host molecule (receptor, enzyme, etc.) which is a manifestation of the conformational requirements imposed by the host. R_{bio} is also represented as a volume in the ligand’s conformation space, engulfing the generalized set of conformations that can, in principle but not necessarily in practice, fit into the binding site. For each host molecule there is a different region of bioactivity reflecting a different set of conformational requirements from the ligand. Cast in these terms, the binding affinity (conformational part) is determined by the percent of overlap between the ligand’s “occupied volume in conformation space” and the host’s pre-prescribed region of bioactivity; i.e.,

$$\text{binding affinity} \propto \%\ \text{overlap} = \frac{V_{\mathrm{conf}} \cap R_{\mathrm{bio}}}{V_{\mathrm{conf}}} \qquad (1)$$

The overlap should be weighted by the Boltzmann factor.

With these definitions, the effect of conformation constraints on bioactivity can be readily discussed. In general, conformation constraints and chemical modifications can change (typically reduce) the flexibility of the molecule and/or introduce structural stereochemical changes (e.g., they may force the molecule into a twisted shape). These, in turn, affect the size and shape of the volume in conformation space V_{conf} occupied by the molecule. Reduced flexibility will be manifested as a decrease in the “occupied volume” V_{conf} (fewer conformations are accessible to the molecules), while the strereochemical effect will be manifested as a shift of the volume $\mathrm{V}_{\mathrm{conf}}$ relative to its original location (a different set of conformations is now accessible to the molecule). Such changes in V_{conf} affect its overlap with the host’s region of bioactivity, R_{bio}, thus changing the ligands binding affinity (eq 1). Based on these concepts, at least three characteristic scenarios for the effect of conformation constraints on bioactivity can be outlined. Of course, in realistic systems a combination of these scenarios is to be expected. Also recall that these scenarios account only for the conformational aspect of the binding affinity (docking). They do not address issues concerning chemical compatibility. The three scenarios are as follows:

(1) *Bioactivity is related to a decrease in the occupied conformation volume* (Figure 1a). A conformation constraint, such as cyclization, often reduces the flexibility of the molecule. This means that the volume in conformation space accessible to the constraint analogue V'_{conf} is smaller than the original conformation volume V_{conf} occupied by the unconstrained molecule. As illustrated in Figure 1a, the percent of overlap (eq 1) depends of the actual conformation volume that remains occupied by the constrained analogue and can vary from 100% to 0%. If the reduced conformation volume V'_{conf} of the constrained analogue falls completely within the region of bioactivity R_{bio}, all of its conformations fit the binding site and its binding affinity is very high. On the other hand, if the reduced volume V'_{conf} falls completely outside the region of bioactivity R_{bio}, none of the conformations fit the binding site and its binding affinity will be zero.

(2) *Bioactivity is related to partially overlapping conformation volumes* (Figure 1b). Chemical modifications, such as point mutations in peptides, may result in series of analogous molecules, all of which exhibit a similar level of flexibility (i.e., occupy similar volumes in conformation volume). In such cases the main effect of the constraints is to shift the new conformation volume V'_{conf} relative to the original conformation volume V_{conf} of the native analogue. The relative bioactivity of the different analogues will depend on the percent of overlap between the different conformation volumes V'_{conf} and the region of bioactivity R_{bio} pre-prescribed by the host. A gradual change in binding affinity is expected across the series of analogues as the percent of overlap changes from one analogue to another. Technically, because the region of bioactivity is rarely known, it can be approximated by the conformation volume occupied by the most potent analogue in the series (which has a maximal overlap with the region of bioactivity).

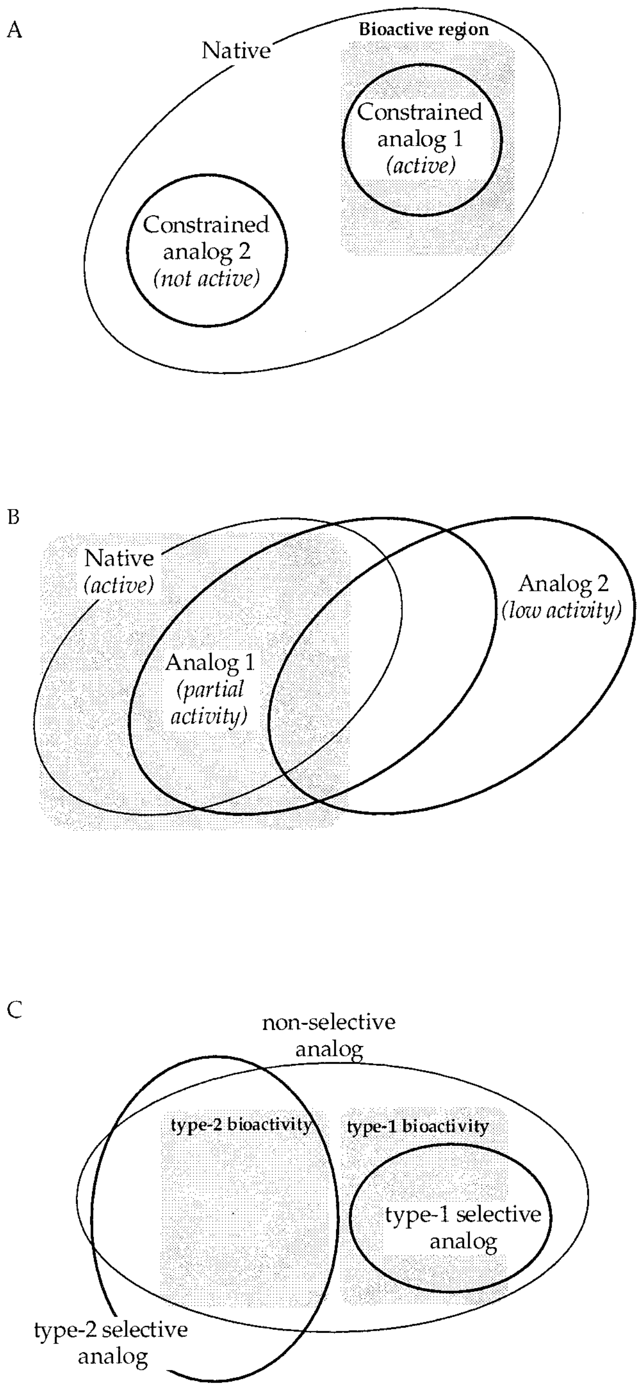

Figure 1. Schematic representation of the effect of flexibility and conformation constraints on the bioactivity of flexible molecules. The ellipsoids indicate the conformation space accessible to each molecular analogue, V_{conf}, while the shaded areas indicate the host-prescribed region of bioactivity, R_{bio}, which includes all possible conformations that can bind to the host (e.g., enzymes or receptors). Three scenarios are illustrated (see text): (A) Different analogues have different conformation volumes; (B) Different analogues have partially overlapping conformation spaces; and (C) Specificity is related to separation and molecular specificity.

(3) *Binding specificity and conformation spaces* (Figure 1c). It is known that conformationally constrained drug analogues often exhibit different selectivity properties when tested on different receptors (or even receptor subtypes). In fact, the prospect of obtaining a selective drug is one of the main reasons why conformationally constrained analogues are studied in the first place. The notion of conformation volumes can be used to explain this phenomenon too (again, only its conformational aspect). Different receptors (or receptor subtypes) have different conformational requirements from the ligand, i.e., they define different regions of bioactivity R'_{bio} and R''_{bio}. A nonselective ligand is flexible enough so that the region conformation space accessible to it V_{conf} overlaps (at least partially) both regions of bioactivity. Namely, it can adopt conformations compatible with either receptor. A selective drug analogue, on the other hand, is characterized by a conformation space volume V_{conf} that preferentially overlaps only one of the two regions of bioactivity: i.e.,

$$(V_{conf} \cap R'_{bio}) \gg (V_{conf} \cap R''_{bio}) \approx 0 \qquad (2)$$

A nonactive analogue will have no overlap with either region of bioactivity. As illustrated in Figure 1c, such a preference should be reflected as a spatial separation of the accessible volumes in conformation space.

Quantifying "Conformation Space"

The above discussion indicated that the molecular conformation space is a useful conceptual framework for addressing questions about molecular flexibility and its relation to binding affinity. However, only recently have these abstract concepts become computationally tractable. The main difficulty associated with quantifying molecular conformation spaces is their high dimensionality. Since every atom is defined by three Cartesian coordinates (x, y, and z), $3N$ coordinates are required to specify a conformation of N atom molecules. Thus, the space that represents the conformations of such a molecule is $3N$-dimensional. This means that even the conformation space of a relatively small polypeptide is extremely high dimensional (hundreds to thousands of dimensions). Luckily, in practice, a much smaller number of dimensions (i.e., coordinates) are sufficient to characterize the essential conformational properties of peptides. For example, the useful ϕ, ψ backbone dihedral angle description reduces the effective dimensionality by an average factor of 10 or more. Further reduction of dimensionality can be obtained by using principal component analysis, which picks out the few most important coordinates required to characterize the conformational diversity of the molecule.[8-13] Recently, we showed, for a range of peptides, that conformation spaces can often be quite accurately represented by as few as three or four principal axes.[14] These projections allow us to construct quantitative energy landscapes for peptides of different lengths.[14-16]

The ability to quantify and visualize molecular conformational spaces, offered by such projection techniques, allows one to quantitatively rationalize the effect of flexibility and conformation constraints on bioactivity.

Methods and Model Systems

Model Systems. The conformation spaces of three polypeptides were analyzed in this study. Two of the three, substance P and RGD containing peptides, are actual drugs of major pharmaceutical importance.

(1) Substance P Analogues. Substance P (SP) is an 11 amino acid neuropeptide of the sequence H$-$Arg$^1-$Pro$^2-$Lys$^3-$Pro$^4-$Gln$^5-$Gln$^6-$Phe$^7-$Phe$^8-$Gly$^9-$Leu$^{10}-$Met$^{11}-$NH$_2$. It belongs to the tachykinin family and is involved, as a neurotransmitter, in a variety of biological activities. Extensive studies showed that the C-terminal half of the molecule, starting at Gln^6, dominates the binding of substance P to the NK1 receptor. For example, Cascieri et al.[17] showed that the SP derivative Gln$^6-$

TABLE 1: The Seven Substance P Analogues Studied[a]

	Arg	Pro	Lys	Pro	Gln	Gln	Phe	Phe	Gly	Leu	Met	IC$_{50}$	−logIC$_{50}$
SP$_{1}$	−	−	−	−	−	−	−	−	−	−	−	1	0.00
SP$_{6}$	−	−	−	−	D	−	−	−	−	−	−	2	−0.30
SP$_{7}$	−	−	−	−	−	D	−	−	−	−	−	250	−2.40
SP$_{9}$	−	−	−	−	−	−	−	D	−	−	−	200	−2.30
SP$_{39}$	−	−	−	−	D	−	−	D	−	−	−	350	−2.54
SP$_{42}$	−	−	−	−	−	D	−	D	−	−	−	>10000	<−4.00
SP$_{122}$	−	−	−	−	D	D	−	D	−	−	−	>10000	<−4.00

[a] L to D substituted amino acids are marked with D. Binding affinities to the NK1 receptor (IC$_{50}$ values in nM) and analogue notation are from Wang et al.[2]

TABLE 2: The Four Arg−Gly−Asp Containing Analogs Studied[a]

analog[b]	FN affinity	VN affinity
Gly−Arg−Gly−Asp−Ser−Pro−Cys	1	1
Gly−**dArg**−Gly−Asp−Ser−Pro−Cys	1	1.1
Gly−Arg−Gly−**dAsp**−Ser−Pro−Cys	0	0
Pen−Gly−Arg−Gly−Asp−Ser−Pro−Cys[c]	0	10

[a] The relative affinities to the fibronectin receptor (FN) and to the vitronectin receptor (VN) are from Pierschbacher and Ruoslahti.[1] [b] dArg and dAsp indicate D-amino acid enantiomers. [c] In the simulations Cys replaced the penicillamine group (Pen) used in the experiments to form the disulfide cyclization.

Met^{11} is much more bioactive than shorter analogues. In a very detailed study, Wang et al.[2] synthesized 512 SP analogues using a systematic D-amino acid replacement strategy (Met^{11} was kept as an L-amino acid and Gly^{9} is achiral). The binding affinity of each analogue to the NK1 receptor was measured and IC$_{50}$ values obtained (i.e., the concentration required for 50% inhibition on NK1). The natural all-L SP peptide was found to have the highest binding affinity relative to all the enantiomeric analogues. Based on this study we selected a set of seven SP analogues for the present analysis. The analogues, detailed in Table 1, were selected to cover a broad range of bioactivity (analogue numbering according to the notation of Wang et al.[2]). Two analogues, SP$_{1}$ (native SP) and SP$_{6}$ were highly bioactive. Three analogues, SP$_{7}$, SP$_{9}$, and SP$_{39}$, showed medium binding affinity. Two analogues, SP$_{42}$ and SP$_{122}$, had very poor binding affinity.

(2) RGD-Containing Peptides. The second group of polypeptides includes septapeptides containing the RGD sequence. The Arg−Gly−Asp (RGD) sequence is the primary recognition site for cell adhesion. This sequence is a probe for cell adhesion of adhesive proteins, such as fibronectin, as well as extracellular matrices.[18] It was found that more than one cell surface receptor exists, and that while they all recognize the Arg−Gly−Asp sequence, these receptors are unique with respect to their individual ligands. For example, one cell surface receptor specifically recognizes fibronectin while another is specific to vitronectin.[19] In a detailed study Pierschbacher and Ruoslahti checked the binding and selectivity of several conformationally constrained Arg−Gly−Asp containing analogues by their inhibition of cell attachment to fibronectin and vitronectin.[1] These researchers concluded that the stereochemistry of the Arg−Gly−Asp sequence itself, as influenced by the enantiomeric substitution of one of its residues or one of its neighboring residues, has a significant influence on selectivity. In the present study we selected four of the analogues, based on the Gly^{1}−Arg^{2}−Gly^{3}−Asp^{4}−Ser^{5}−Pro^{6}−Cys^{7} sequence, studied by Pierschbacher and Ruoslahti. Table 2 details the four analogues and their relative binding affinities to the fibronectin receptor (FN) and to the vitronectin receptor (VN). These analogues reflect a broad range of bioactivity. The all L-amino acid peptide as well as the D-Arg^{2} substituted analogue exhibit a similar affinity to both receptors. The D-Asp^{4} analogue lost its binding affinity altogether, while the cyclic analogue was very selective; its affinity to the vitronectin receptor was 10-fold greater than that of the native peptide, while its affinity to the fibronectin receptor was negligible. In the present study the cyclic analogue was generated by substituting Gly^{1} with Cys and forming a disulfide bridge with Cys^{7} (in the original study a penicillamine group attached to Gly^{1} was used to form the disulfide bridge).

(3) Alanine Hexapeptide Analogues. The third peptide family analyzed in this paper includes conformationally constrained analogues of the alanine hexapeptide. Four hexapeptide analogues were studied: unconstrained linear (Ala)$_{6}$, backbone cyclic (Ala)$_{6}$, and two Ala to Pro substitutions: (Ala)$_{2}$−Pro−(Ala)$_{3}$ and (Ala)$_{2}$−(Pro)$_{2}$−(Ala)$_{2}$. Both cyclization and Pro substitutions are expected to reduce the flexibility of the molecule. In a previous study, using the topological mapping methodology,[20] we showed that the energy landscape of linear (Ala)$_{6}$ was very different from the energy landscape of its backbone cyclic analogue.[21] The two landscapes differed in their internal connectivity, the range of energies represented and the surface roughness. Therefore, despite the lack of specific bioactivity, this peptide is a good model for studying the effect of cyclization on the size of the molecular conformation space.

Computational Methods. The analysis procedure used in this study follows three steps. First, a large conformation sample is constructed. Then principal component analysis is used to project the sampled conformation space onto a small number of principal directions. Finally, the weighted multidimensional overlaps of conformation spaces are calculated.

(1) *Conformation Sampling*. Performing a conformational ensemble sampling of each peptide is necessary for representing the molecule's conformation space. In principle, one wants as complete as possible representation of this space. However, because of the large volume of conformation space available to polypeptides, sampling has to be used instead of a systematic conformational search. A variety of sampling approaches are available.[22] The procedure used in this study to sample the conformation space of the above peptides was previously reported.[21] Briefly stated, for each peptide a sample of 500 conformations is collected from a 500 ps high temperature molecular dynamics trajectory, simulated at 1000 K (it was shown that there are no cis/trans transitions of the peptide bond at this sampling temperature[23]). Each high temperature conformation was then gradually cooled to 300 K, after which it was minimized to the nearest local minimum. The initial structures for all noncyclic peptides were the extended conformations; for cyclic peptides randomly selected cyclic conformation were used. All simulations were performed with the molecular dynamics program CHARMM[24] and the CHARMM all-atom force field,[25] using 2 fs time steps, 15 Å cutoffs, SHAKE constraints on bonds to hydrogen atoms, and a distance-dependent dielectric constant.

(2) *Principal Component Projections*. Principal component analysis (PCA) projects multidimensional data onto low-dimensional subspaces.[12] If the distribution of the multidimensional data is nonisotropic, PCA will identify a low-dimensional subspace that best describes it. Namely, it selects a new (minimal) set of principal axes that best preserve the distances between the conformations, enabling visualization of the spatial relations between the data points. One of the advantages of PCA is that the normalized eigenvalues λ_i, associated with each principal axis (eigenvectors), are directly related to the effective dimensionality of the projection and to the average error associated with it. Principal axes are sorted according to their normalized eigenvalues λ_i. The larger the eigenvalue the more efficient is the projection onto that axis (reflecting a large variance for the data in that 1D projection). In recent years PCA has become a common method for analyzing complex molecular data. Applications include analysis of molecular dynamics trajectories,[9,13] conformation sampling,[10,11] and conformation clustering.[26] Using a variant of this method, named principal coordinate analysis (PCoorA),[8] Becker and collaborators have generated quantitative 3D maps of the energy landscapes of peptides.[15,16] In this study we apply PCoorA to project the multidimensional conformation samples onto 2, 3, or 4 dimensional subspaces. It was shown elsewhere, at least for the Arg−Gly−Asp containing peptides, that the principal 2 and 3 dimensional subspaces represent the multidimensional data to accuracy greater than 70%.[14]

The details of PcoorA were discussed elsewhere.[14,26] For the present application it should be noted that when conformations of two analogous molecules are to be compared they must be projected together onto the *same* subspace (starting from a joint distance matrix). As a consequence, the distance measure used should be based on features common to the two molecules. In this study we use two distance measures based the conformation of the peptide backbone, which is common throughout each family of peptide analogues. The first is the root-mean-square distance (RMSD) in Cartesian coordinates and the second is in the peptide's backbone dihedral angle space (ϕ, ψ).

(3) *Weighted Multidimensional Overlaps of Conformation Spaces*. In the following discussion multidimensional overlaps between the volume in conformation space occupied by one molecule and those occupied by another are calculated (based on joint principal projections). Because conformations are not uniformly distributed throughout the available region in conformation space (they are weighted according to energy by the Boltzmann factor), a simple geometric overlap is not enough. Rather, the calculated overlaps should be weighted by the observed population distribution. To overcome the sparsity of the data, a multidimensional grid, defined in principal coordinates, is used. Each relevant principal axis is divided into segments, and the number of conformations that lay within each multidimensional cell (bin) is counted. Because the spread of the data points is different from one principal axis to the other, the number of grid segments along each axis is set so that the information contents in the multidimensional bins are roughly equivalent. Thus, axes with larger eigenvalues are divided into more bins relative to axes with small eigenvalues which are divided into fewer bins.

Following the assumption that the accessible region in conformation space is continuous for each molecule, a smoothing algorithm is applied to the above multidimensional grid (typically 3D or 4D grids). The goal of this algorithm is to smooth "holes" in the distribution that may be caused by insufficient sampling and/or by the discretization of the continuous space. Holes in an otherwise populated region would be filled according to the population of their neighbors (nearest neighbors have the weight 1/2 and second nearest neighbors are weighted by 1/3). As a result of the smoothing, each grid point carries a noninteger weight, reflecting the relative population in that region in conformation space. These population factors are denoted as $P_k{}^{(i)}$, where k is the index of the multidimensional grid point and (i) is the index of the molecule. Overlap between the regions in conformation space occupied by molecules i and j is given by the following expression:

$$\text{overlap} = \sum_k P_k^{(i)} P_k^{(j)} \tag{3}$$

where the summation is over the grid points k. In fact, two overlap measures can be defined. The two differ with regard to which grid points are included in the summation. In the first overlap measure, denoted $O_\&$, grid points are included in the summation only if the population factors at these grid points P_k, for both molecules, are larger than some threshold ϵ,

$$O_\& = \frac{1}{M} \sum_k P_k^{(i)} P_k^{(j)} \quad (k|P_k^{(i)} > \epsilon \text{ and } P_k^{(j)} > \epsilon) \tag{4}$$

The summation is divided by the total number M of grid points included in the summation in order to obtain an overlap density (the amount of overlap per unit volume in conformation space). In the second overlap measure, denoted $O_{||}$, a grid point is included in the summation if the population factor P_k at this grid point is greater than the threshold ϵ for at least one of the two molecules,

$$O_{||} = \frac{1}{M} \sum_k P_k^{(i)} P_k^{(j)} \quad (k|P_k^{(i)} > \epsilon \text{ or } P_k^{(j)} > \epsilon) \tag{5}$$

We found that the first overlap measure $O_\&$ is too restrictive. The requirement that the grid point P_k for both molecules will be higher than the threshold ϵ makes this measure very sensitive to the precise definition and placing of the grid. Much more stable results were obtained for the less restrictive overlap measure $O_{||}$, rendering it a more useful quantity. The role of the threshold value ϵ will be discussed below.

Results

The above analysis, i.e., conformation sampling followed by joint principal component projections onto low-dimensional subspaces, was applied to the three groups of peptides described above. The projections are discussed in terms of the scenarios suggested in Figure 1, and the relation between bioactivity and conformation space is pointed out.

Reduction of Conformation Space. Cyclization is probably the most constraining modification applied to bioactive peptides in an attempt to improve their potency. Clearly, cyclic analogues are much less flexible than their linear counterparts. In the language of conformation space, cyclic analogues are expected to occupy a much smaller conformation volume compared to the unconstrained molecules. Nonetheless, questions such as to what extent does a cyclization reduce the available conformation space, and whether the reduced space is a subset of the native conformation space, are for the most part a matter of speculation. The analysis of the two hexapeptide analogues, linear (Ala)$_6$ and cyclic (Ala)$_6$, allow us to offer quantitative answers to these puzzles.

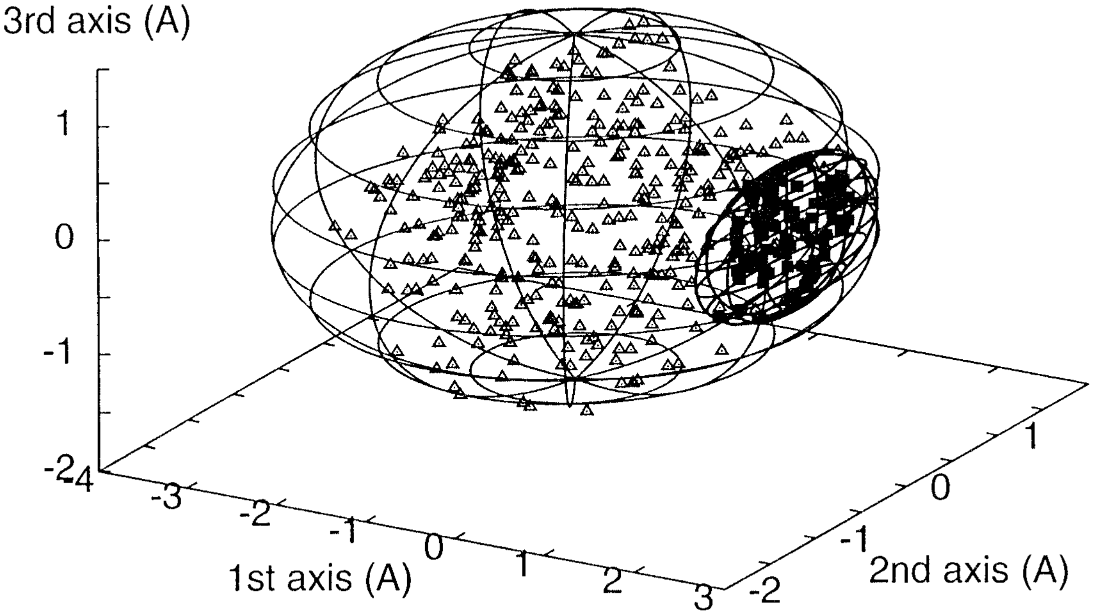

Figure 2. Joint projection of the available conformation spaces of linear $(Ala)_6$ (triangles) and the cyclic $(Ala)_6$ analogue (filled squares) onto the optimal 3D principal axes (see text). The symbols indicate the projected conformations and the ellipsoids engulf the volume occupied by the projected points. This projection shows that the conformation volume accessible to the cyclic analogue is a small subset of the conformation volume accessible to the linear peptide, amounting in this case to 12% of the original volume. This reduction reflects the loss of flexibility and conformational entropy upon cyclization.

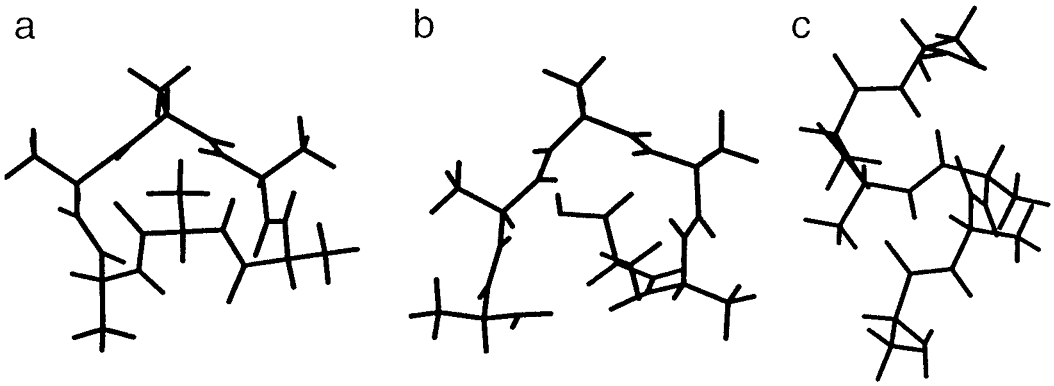

Figure 3. Conformation similarity in joint PCA projections. The structurally similar conformations, the cyclic-$(Ala)_6$ conformation (a) and the linear $(Ala)_6$ conformation (b), are neighbors in the joint projection of Figure 2 (both taken from the overlap region). The third conformation (c) represents a dissimilar helical conformation (the lowest energy conformation of linear $(Ala)_6$), which in the projection appears far from the overlap region in Figure 2.

As described above, 500 conformations were sampled for each peptide analogue. The 50 highest energy conformations from each set were removed, resulting in 450 conformations for each peptide. A joint projection of the two peptides was performed based on the 900 × 900 joint distance matrix. Backbone rms distances in Cartesian space were used to measure the distances between conformations in the data set. Figure 2 shows the joint projection of the two analogues, linear $(Ala)_6$ and cyclic $(Ala)_6$, onto the optimal 3D subspace defined by the first three principal axes (i.e., the three principal eigenvectors associated with the largest eigenvalues). The normalized eigenvalues associated with these axes are 22.5%, 13.5%, and 9% (the next three principal axes carry much less information, as indicated by their smaller normalized eigenvalues which are 6%, 6%, 4%). Namely, in this representation individual distances between points are accurate, on the average, only to about 50%. However, detailed analysis of the distribution of errors in PCA projections has shown that for practical purposes the actual quality of the projection is much higher than that. It was shown that a relatively small number of poorly represented points skew the average error to larger values, and that the majority of the distances are represented to a much better accuracy (often the median error is smaller by a factor of 2 compared to the average error).[14] This means that, for the most part, the error in the projection is only on the order of 25%. Figure 3 demonstrates that the joint projection indeed reflects conformational similarity. Figures 3a and 3b show two conformations, one of the cyclic analogue and the other of the linear analogue, which are neighbors in the joint projection (taken from the region of overlap between the two conformation volumes). The structural similarity between these two conformations is apparent. Figure 3c, on the other hand, shows the helical structure of the lowest energy conformation of $(Ala)_6$. In the joint projection this conformation, which is very different from the first two, appears quite far away from the overlap region.

Figure 2 shows that, as expected, the conformation space available to the cyclic analogue V'_{conf} is dramatically smaller than the conformation volume, V_{conf}, occupied by the linear analogue, reflecting its reduced flexibility. Furthermore, at least in this case, we find that the space accessible to the constrained

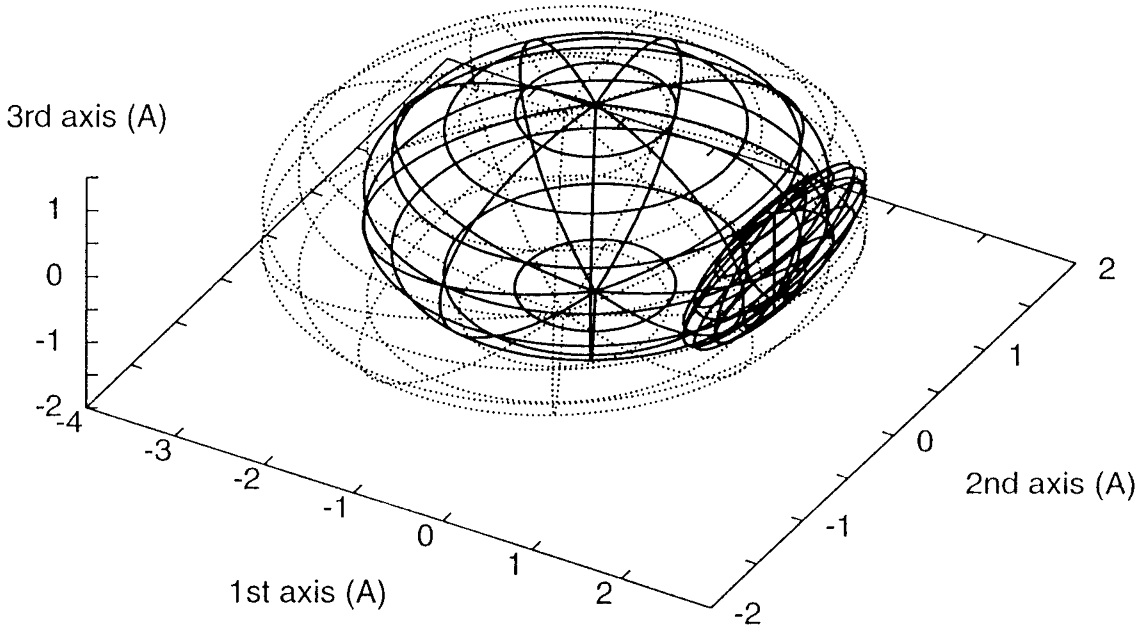

Figure 4. Projection of the available conformation spaces of three alanine hexapeptide analogues onto a joint principal 3D subspace (for clarity, only the ellipsoids engulfing the occupied volumes are shown). The projection includes the two analogues shown in Figure 2, linear $(Ala)_6$ (dashed line) and the cyclic $(Ala)_6$ analogue (bold line), as well as the $(Ala)_2-(Pro)_2-(Ala)_2$ analogue (solid line). Both cyclization and the double Pro substitution reduce the flexibility of the molecule. However, the available conformation volume of $(Ala)_2-(Pro)_2-(Ala)_2$ is reduced only to approximately 55% of the original size, compared to 12% for the cyclic analogue. It is also seen that the two different constraints restrict the molecule to different parts of the original conformation space (similar to the scenario depicted in Figure 1a).

analogue is indeed a subset of the original unconstrained conformation space ($V_{\mathrm{conf}} \supset V'_{\mathrm{conf}}$). To quantify this effect the conformation volume of each analogue is estimated by the volume of a 3D ellipsoid, V^{3D}, which engulfs all of the points associated with that conformation sample in the principal 3D subspace. The ellipsoids are calculated by diagonalizing the 3 $\times$ 3 covarience matrices of the principal axes.[15] In the present case we find that the principal 3D volume of the cyclic analogue is only 12% of the conformation volume available to the native unconstrained molecule. It can be argued that the reduction in conformation volume as a result of the cyclization reflects a similar decrease in conformational entropy (where entropy is proportional to the logarithm of the conformation volume). Since the contribution of the higher principal axes is decreasingly small, the logarithm of the 3D volume, $\ln V^{3D}$, is a very rough estimate of the relative conformational entropy. Although this is only a rough estimate, in the present case the ratio of the two logarithms is 2.5, indicating that the molecule loses about 60% of its conformational entropy upon cyclization. If this peptide had any biological activity, the above result would be a manifestation of the first scenario schematically depicted in Figure 1a. Whether the cyclic analogue is active or not depends on the overlap between the (small) volume it occupies in conformation space and the region of bioactivity R_{bio} prescribed by the host.

Figure 4 shows the outlines of the available conformation spaces of three $(Ala)_6$ analogues jointly projected onto the same principal 3D subspace (the nominal accuracy of this 3D projection is 48%). In addition to the two analogues shown in Figure 2, linear $(Ala)_6$ and cyclic $(Ala)_6$, Figure 4 also includes the conformation space of the double proline substituted analogue $(Ala)_2-(Pro)_2-(Ala)_2$. The triple projection shows that the double Pro substitution also decreases the flexibility of the molecule, as demonstrated by a reduction in the volume it occupies in conformation space. Note, however, that the conformation constraint introduced by the double Pro substitution restricts the molecule to a different part of its original conformation space. This situation demonstrates the scenario depicted in Figure 1a, i.e., that different constrained analogues may occupy different parts of the original conformation space, resulting in different bioactivities. Applying the above 3D volume calculation we find that the volume in conformation space occupied by the double Pro substituted analogue $(Ala)_2-(Pro)_2-(Ala)_2$ is only 55% of the volume available to the native peptide (compared to 12% available to the cyclic analog). Repeating the same calculation with the single Pro substituted analogue $(Ala)_2-Pro-(Ala)_3$ resulted in a much smaller reduction in the available space. For this analogue the available volume in conformation space was 88% of that available to the native peptide, indicating that a single Pro mutation has a much smaller effect on the peptide's flexibility. To conclude, these results, which are summarized in Table 3, indicate that the relative size of the conformation volumes occupied by the above four peptide analogues is $(Ala)_6 > (Ala)_2-Pro-(Ala)_3 > (Ala)_2-(Pro)_2-(Ala)_2 >$ cyclic-$(Ala)_6$.

TABLE 3: Relative Conformation Volumes of Four Analine Hexapeptide Analogues

peptide	% overlap with $(Ala)_6$
$(Ala)_6$	100%
$(Ala)_2-Pro-(Ala)_3$	88%
$(Ala)_2-(Pro)_2-(Ala)_2$	55%
cyclic-$(Ala)_6$	12%

Partially Overlapping Conformation Spaces. A nice demonstration of the second scenario, in which bioactivity is related to the degree of overlap between conformation spaces, was found in the family of substance P (SP) analogues studied.

The seven SP analogues specified in Table 2, which differ from one another by up to three enantiomeric L to D amino acid substitutions, were subjected to the same conformation sampling protocol described above. This resulted in a 3500 conformation sample. Because the range of energies spanned by each set of 500 conformations was very broad (between 70 and 80 kcal/mol) the subsequent conformation space analysis was applied only to the 100 lowest-energy conformations of each peptide. The energy range spanned by each set of 100 conformations was only 16 to 28 kcal/mol. This restriction of the conformation sample is equivalent to a rough application of the Boltzmann factor, which preferentially weights the low energy structures. Previous studies have shown that bioactivity of SP (i.e., binding to the NK1 receptor) is dominated by the C-terminal half of the molecule, starting with residue 6.[2,17] To highlight the role of flexibility and conformation constraints in the bioactive part of the molecule, the distance measure used to compare peptide conformation was the Cartesian backbone rms distance of the C-terminal residues 6 to 11. When compared to distances based on the whole 11-residue backbone it was found that the contribution of the N terminal residues to the total rms distance is relatively small (28% of the total distance).

The functional difference between the two halves of the peptide is strongly correlated to the molecule's flexibility patterns: the functional C-terminal was found to be significantly more flexible than the functionless N-terminal. Figure 5 shows three aligned overlays of the 20 lowest energy conformations of native SP. Even within this small sample the flexibility of this molecule is striking, and so is the separate clustering of its two termini. Aligning either the five N-terminal residues (Figure 5a) or the six C-terminal residues (Figure 5b) results in a complete misalignment of the other terminal. It is interesting to note that the residues in the functionless N-terminal adopt very similar conformations (the Cα rms distance of these five residues is 0.64 Å). On the other hand, even when optimally aligned, the functional C-terminal residues exhibit a broad range of conformations (the Cα rms distance of these six residues is 2.46 Å). The overall Cα rms distance of these 20 conformations, when aligned according to all 11 Cα atoms, is 3.12 Å (Figure 5c). This result supports the use of a distance measure based on the C-terminal residue. It also demonstrates that the molecule's flexibility is adequately represented by its 100 lowest energy conformations.

Using the above distance measure, based on the C-terminal residues, a 700 × 700 joint distance matrix for all seven peptides was constructed and submitted to PCoorA. Because all seven conformations are rather similar, the resulting projection was of a lower quality compared to projections obtained for other systems. The accuracy of the best four-dimensional (4D) projection was only 40% (the contributions of the individual axes were: 20.2%, 8.0%, 5.9%, and 5.4%). The contribution of the remaining individual axes, however, was even smaller. Individual axes from the eighth principal axis and on contributed less than 2% to the overall accuracy (less than 1% from the 15th axis and on). A better accuracy, close to 50%, was obtained when the 4D joint projection was based on the lowest 50 conformations for each of the seven peptides (the contributions of the individual axes were: 25.4%, 9.6%, 7.0%, and 6.1%). Despite the rather poor average quality of the 4D projections, the fact that the contribution of the other principal coordinates is diminishingly small indicates that these 4D projections capture most of the anisotropy in the system. Looking at the projections, we found that all seven molecules occupy conformation volumes of comparable sizes (i.e., they exhibit similar flexibility).

Next, the weighted 4D overlaps, between the conformation spaces V_{conf} of the seven SP analogues and the NK1 receptor's region of bioactivity R_{bio} should be calculated. Because this region of bioactivity is not known in itself, it is approximated by the conformation volume occupied by the most potent analogue in the series $V_{\mathrm{conf}}^{\mathrm{potent}}$, which has a maximal amount of overlap with the receptor's region of bioactivity. Namely, $R_{\mathrm{bio}} \approx V_{\mathrm{conf}}^{\mathrm{potent}}$, and eq 1 is rewritten as

$$\text{binding affinity} \propto \text{\% overlap} = \frac{V_{\mathrm{conf}} \cap V_{\mathrm{conf}}^{\mathrm{potent}}}{V_{\mathrm{conf}}} \qquad (6)$$

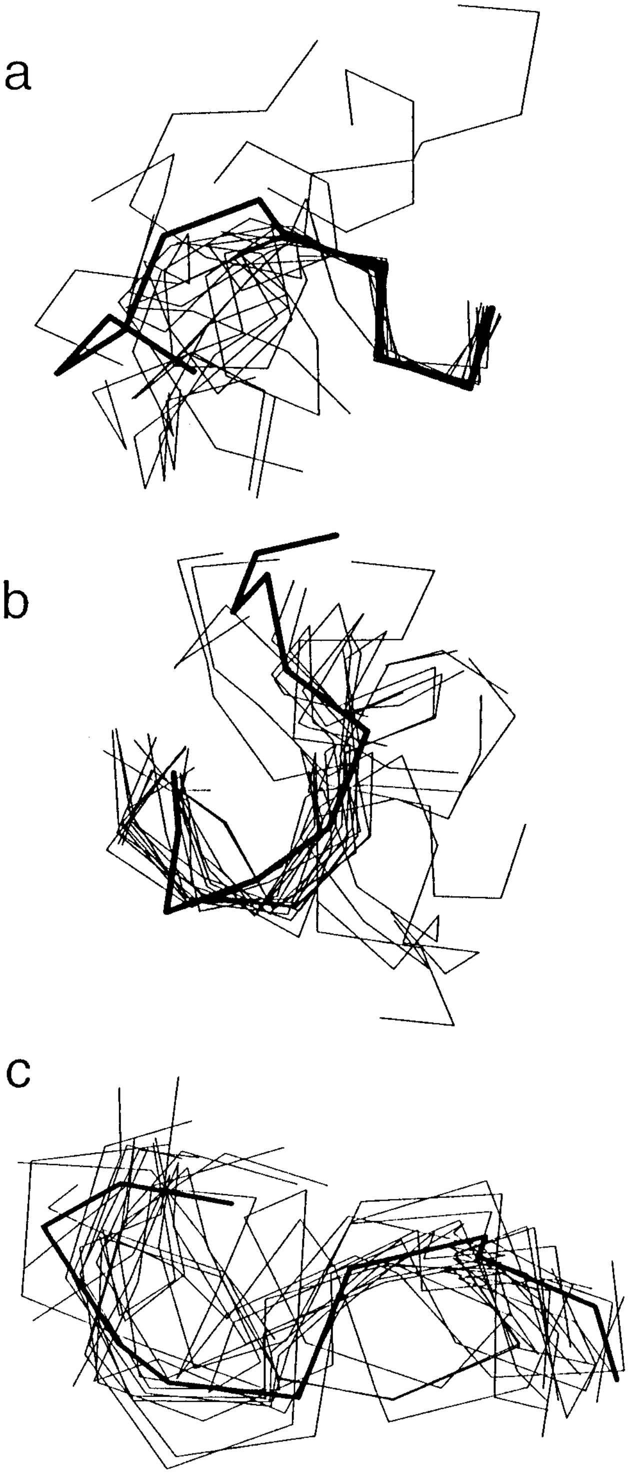

Figure 5. Three aligned overlays of the 20 lowest-energy conformations of native substance P. The lowest energy conformation appears in bold: (a) alignment based on the Cα atoms of the 5 (functionless) N-terminal residues; (b) alignment based on the Cα atoms of the 6 (functional) C-terminal residues; (c) alignment based on the Cα atoms of all 11 residues. The flexibility of this molecule, the separate clustering of the two termini and the fact that the functional C-terminal is significantly more flexible than the functionless N-terminal, is clearly seen.

In this case the native SP has the largest binding affinity and the weighted 4D overlaps are calculated between the conformation volumes of the six constrained SP analogues and the conformation volume of native SP. As discussed above, the calculations of the 4D overlaps are performed with a binning algorithm in which the four axes are divided nonhomogeneously to create equal weight bins. Axes with larger eigenvalues are divided into more segments than axes with small eigenvalues. The number of segments is set in proportion to the ratio between eigenvalues. The ratio between the four largest eigenvalues in the joint 4D projection of the SP analogues is roughly 4:1.5:1:1 (for both the 50 conformation and 100 conformation samples). Thus the number of divisions applied to these four axes should be 8, 3, 2, and 2, respectively. This division, however, results in 96 4D bins, which is too large for the size of the data. To overcome this problem, fewer divisions along the first and second principal axes are taken, resulting in a 4:2:2:2 binning scheme. Namely, the first principal axis is divided into 4 segments, the second axis into 2 segments, and the third and fourth axes into 2 segments each. This partition results in 36 4D bins. An alternative 4:3:2:2 binning scheme (with 48 4D bins) was also tested. Both binning schemes yielded good results.

The 4D overlaps, between the 4D volumes in conformation space occupied by each of the six SP analogues and the volume occupied by the native SP are calculated using the less restrictive "weighted overlap measure" $O_{||}$ defined in eq 5. Since we are interested in the percent of overlap between an SP analogue and the native SP, the calculated $O_{||}$ value is normalized by the self-overlap of native SP with itself. Thus, the overlaps are calculated using the following equation:

$$O_{||} = \frac{1}{O^{\mathrm{SP}}} \frac{1}{M} \sum_k P_k^{(i)} P_k^{\mathrm{SP}} \quad (k \mid P_k^{(i)} > \epsilon \mid\mid P_k^{\mathrm{SP}} > \epsilon) \qquad (7)$$

$P_k^{(i)}$ is the population factor for molecule (i) at the 4D grid point k, P_k^{SP} is the population factor for native SP, M is the number of grid points included in the summation, and O^{SP} is the normalization factor reflecting the "self-overlap" of native SP with itself (i.e., the result obtained when applying eq 5 to calculate the overlap of native SP with itself).

Because the threshold ϵ changes the number M of grid points included in the summation, the results for both overlap measures, $O_{\&}$ and $O_{||}$, depend on this value. In general, the role of the threshold ϵ is to control the effect of the hole filling procedure, which causes the population factors $P_k^{(i)}$ of "filled holes" to increase from zero to a small noninteger number on the order of 1. To study the effect of the threshold on the calculated overlaps and to select the most appropriate threshold value, the calculation was repeated with a series of threshold values, $\epsilon = 0, 0.25, 0.5, ..., 1.5$. Fortunately, in all cases we were able to find a range of threshold values for which the calculated overlap was not sensitive to small changes in the value of ϵ. Subsequent overlap calculations were restricted to this region of stability. For sample size of 100 conformations per molecule (a total of 700 conformations) the results were stable for threshold values in the range $\epsilon = 1.00-1.25$. For the smaller sample size (50 conformations per molecule) a lower threshold is necessary in order to overcome the space data. The stability region in this case was at threshold values in the range $\epsilon = 0.50-0.75$.

Figure 6 shows the excellent correlation between the experimental bioactivity of the seven SP analogues, measured by $-\log(\mathrm{IC}_{50})$ for the binding affinity to the NK1 receptor[2] (see Table 2) and the calculated percent of overlap, $O_{||}$. Recall that $O_{||}$ measures the overlap between the conformation space volumes occupied by the SP analogues and the volume occupied by native SP (maximal binding affinity). This very strong correlation indicates that the scenario of "partially overlapping conformation spaces" (Figure 1b) suits this family of peptides very well. Evidently, the small conformation constraints, imposed by the enantiomeric substitution of L- to D-amino acids, shift the volume in conformation space occupied by these analogues relative to volume occupied by native SP (which represents the region of bioactivity R_{bio} imposed by the NK1 receptor).

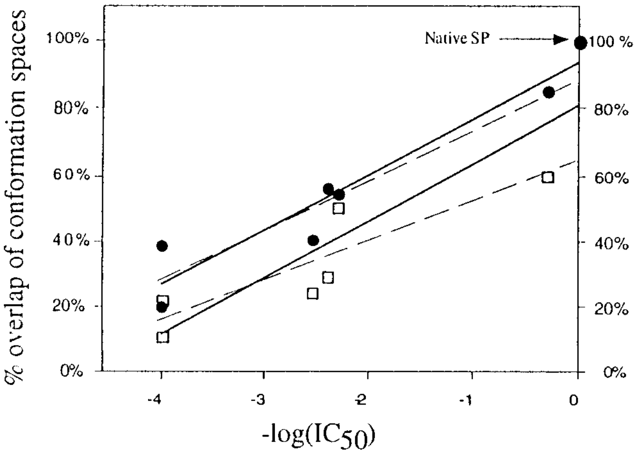

Figure 6. A strong correlation between the experimental bioactivity of the seven conformationally constrained substance P (SP) analogues, measured by $-\log(\mathrm{IC}_{50})$ for the binding affinity to the NK1 receptor (see Table 1), and the calculated percent of overlap, $O_{||}$. Overlap is calculated between the 4D conformation space volumes occupied by these analogues and the 4D conformation space volume occupied by native SP (which exhibits maximal binding affinity and represents the receptor's region of bioactivity R_{bio}). The observed correlation indicates that this family of peptides behaves according to the partially overlapping conformation spaces scenario suggested in Figure 1b. Conformation space overlaps were calculated using two alternative grids: solid circles were calculated on a 4:2:2:2 4D-grid, empty squares were calculated on a 4:3:2:2 4D-grid (see text). Solid lines are linear regression fits to the seven data points (the upper line is calculated for the 4:2:2:2 grid results (solid circles)). Dashed lines are similar linear fits calculated only for six data points, disregarding the "anchor" point of native SP.

The straight lines in Figure 6 are a linear regression fit to the seven data points. An excellent correlation was obtained when the overlaps were calculated using a 4:2:2:2 grid (filled circles; the correlation factor in this case was 0.96). The linear equation correlating the two quantities is

$$-\log(\mathrm{IC}_{50}) = -5.34 + 0.055xO_{||} \qquad (8)$$

where $O_{||}$ is in the range 0% to 100%. A very good correlation between bioactivity and conformation space overlaps was also obtained when the overlap $O_{||}$ was calculated using the alternative grid 4:3:2:2 (empty squares). The linear regression correlation factor in this case was 0.90, but the values were shifted toward lower values (the fitted line crosses the $-\log(\mathrm{IC}_{50}) = 0$ axis at 81%, compared to 94% for the first grid). The linear equation correlating bioactivity and conformation space overlaps calculated with the 4:3:2:2 grid is

$$-\log(\mathrm{IC}_{50}) = -4.22 + 0.047xO_{||} \qquad (9)$$

The dashed curves in Figure 6 are linear fits to the same data calculated after removing the native SP "anchor" point at 100% overlap (only 6 data points to fit). The quality of these fits was

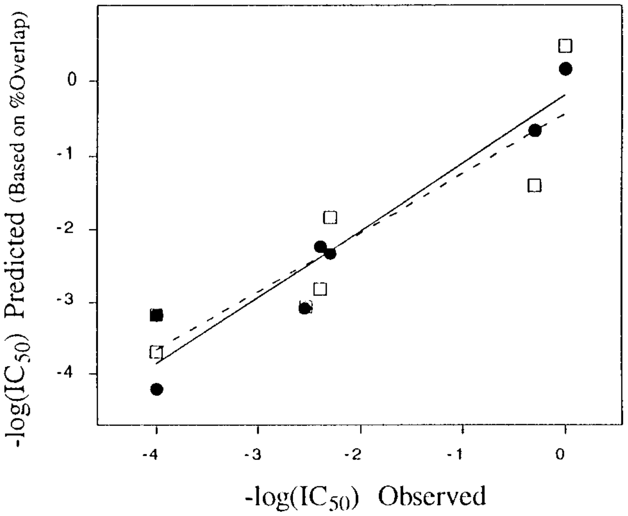

Figure 7. QSAR-type plot comparing the experimentally observed binding affinities of the seven SP analogues (represented as $-\log(IC_{50})$ values) to the predicted binding affinities for this set of molecules. The predicted $-\log(IC_{50})$ values are calculated based on the degree of conformation space overlaps (Figure 6). Filled circles were calculated using the higher quality eq 5 (the linear fit to the data is shown by the solid line). Empty squares were calculated using eq 6, which is of a somewhat lower quality (the linear fit to the data is shown by the dashed line).

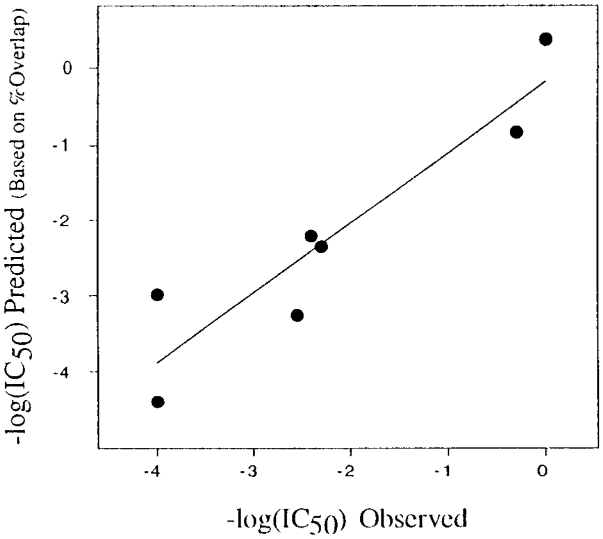

Figure 8. QSAR-type plot similar to Figure 7, but this time each point is predicted using a QSAR equation (similar to eqs 5 and 6) calculated from the other six points, excluding the data point to be predicted (using the 4:2:2:2 grid). The slope of the line is 0.93 and the regression *R* factor 0.927.

similar to those which include the data point representing native SP, although the slopes were slightly different (by 10% for the better results obtained using the 4:2:2:2 grid, and by 30% for the 4:3:2:2 grid).

The observed strong linear correlation suggests that, at least in this case, the percent of overlap in conformation space can be used as a QSAR-type descriptor for predicting the binding affinity of these conformationally constrained analogues. Figure 7 is a QSAR-type plot comparing the experimentally observed binding affinities of the seven peptides (represented as $-\log(IC_{50})$ values) to binding affinities calculated based on the degree of conformation space overlaps. Filled circles were calculated using the higher quality eq 8. The slope of the linear fit to the data (solid line) is 0.92 and the regression *R* factor is 0.96. Empty squares were calculated using eq 9, which is of a somewhat lower quality (the linear fit to the data is shown by the dashed line). It thus seems, that at least for the given data set the overlap between conformation volumes has a good predictive power.

The predictive value of this measure is more justly tested when the predicted point is not part of the training set. Figure 8 shows the same type of comparison between observed bioactivity and predicted bioactivity as in Figure 7, but this time each point is calculated by a QSAR equation (similar to eqs 8 and 9) calculated for the other six points (excluding the data point to be predicted). The slope of the line in Figure 8 is 0.93 and the regression *R* factor 0.927. Figure 8 clearly shows that the correlation is again very strong, indicating that the overlap in conformation space is indeed a reliable QSAR-type descriptor.

Considering the many uncertainties and errors associated with the data points (both experimental and theoretical), the high quality of the correlation is both surprising and reassuring. The fact that very similar predictions were obtained for both grid choices is a tribute to the stability and validity of the observed correlations. The relative insensitivity to small changes in technical parameters indicates that the suggested analysis may be a useful practical approach for quantifying the effect of conformation constraints in the context of QSAR. The observed correlation also suggests that the method is relatively insensitive to the many unavoidable errors in the data. Errors in the experimental data are due to the coarse way in which Wang et al.[2] determined the binding affinities of the conformationally constrained SP analogues (this was done to enable fast screening of the very large number of molecules studied). In particular, it should be noted, that the binding affinities of the inactive analogues were determined only as $IC_{50} > 10\,000$ nM ($-\log(IC_{50}) < -4.0$). This inevitably affects the accuracy of the experimental data at the low-activity ends of Figures 6, 7, and 8. Errors are, of course, also inherent to the theoretically calculated overlaps. These errors originate from incomplete sampling, errors in the PCA projection and sensitivity to the smoothing algorithm.

It should be noted that, in contrast to the good results obtained in the present work, other authors were unsuccessful in fitting a QSAR model to these data. In the original experimental work of Wang et al.[2] the authors tried to construct a QSAR model for the conformationally constrained SP data set, using the Free−Wilson approach,[27] in which the partial contribution from each amino acid is additive and independent of its neighbors (a very strong assumption as far as conformation constraints are concerned). A QSAR model was indeed formulated based on the subset of 189 high and moderate affinity peptides (out of the 512 peptides studied), but the overall fit of the data to that model was only marginal. A similar poor fit to a QSAR model was reported by Eriksson et al.[28] with 39 conformationally constrained SP analogues (these authors used the partial least-squares fit method).

Spatially Separated Conformation Spaces and Specificity. The third scenario, schematically sketched in Figure 1c, relates the concept of conformation spaces to molecular specificity, which is observed when there is more than one receptor (or receptor subtype) that bind the bioactive molecules. The third group of molecules analyzed in this study, the four conformationally constrained Arg−Gly−Asp (RGD) containing peptides, gives a very good example for this scenario.

As discussed above, there are several different receptors that bind the RGD sequence, which is the primary recognition site for cell adhesion. Based on the work of Pierschbacher and Ruoslahti,[1] which showed that conformation constraints vary

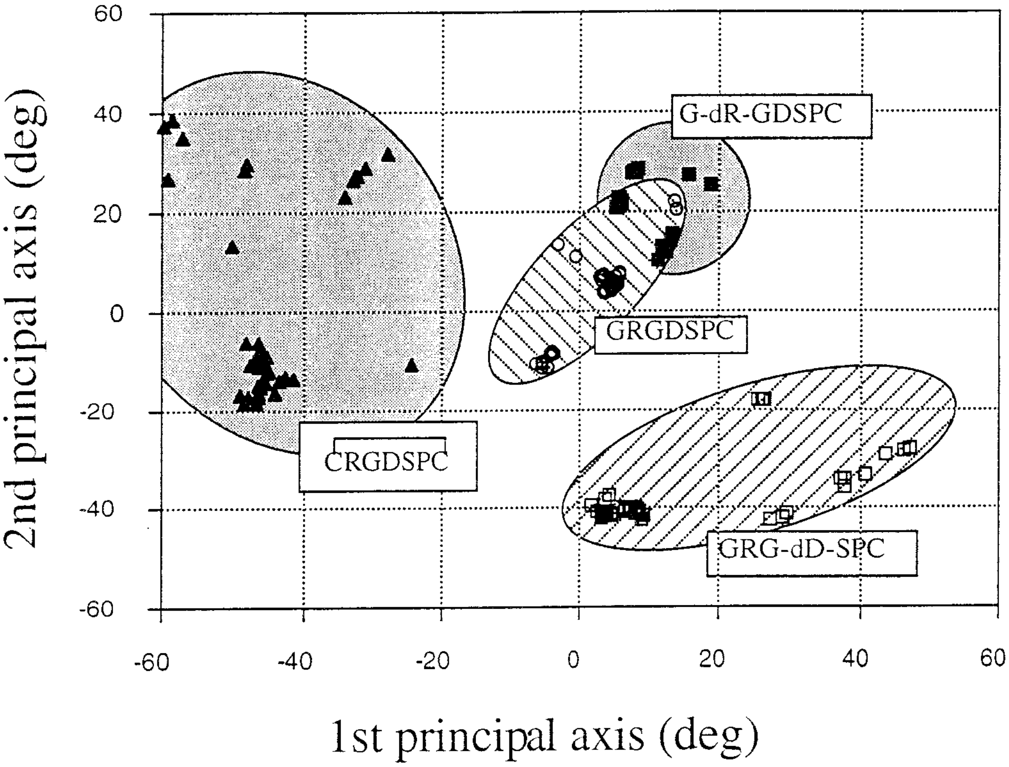

Figure 9. Best 2D projection (first two principal axes) of four conformationally constrained Arg−Gly−Asp containing analogues (details in Table 2). The region of conformation space occupied by each of the peptides is highlighted by a schematic ellipsoid. The notation -dR- and -dD- indicate the D enantiomer of that specific amino acid. The observed separation in conformation space correlates with binding affinities and specificity to receptor subtypes, in agreement with the scenario outlined in Figure 1c. The two nonspecific peptides, GRGDSPC and G-dR-GDSPC, occupy one area in conformation space. The nonactive analogue, GRG-dD-SPC, occupies a different area and so does the potent VN-specific cyclic analogue. The projection is based on dihedral angle distances of seven torsion angles in the RGD region.

the binding affinity and specificity of RGD-containing septapeptides, four such peptides were selected for the present analysis. The four RGD-containing septapeptides selected (Table 2) reflect a broad range of bioactivities, varying in their relative affinities to the fibronectin receptor (FN) and to the vitronectin receptor (VN). Two of the peptides were active but nonselective (the unconstrained native and the peptide with an enantiomeric substitution D-Arg2). One conformationally constrained peptide was not active at all (with an enantiomeric substitution D-Asp4). The fourth peptide, subjected to an end-to-end disulfide cyclization was very selective. It has an extremely high binding affinity to the vitronectin receptor (VN) but practically no affinity to the fibronectin receptor (FN). According to the concepts suggested in Figure 1c, this range of activities and binding specificities indicates that the different peptides occupy disjoint regions in conformation space. The results of the calculation verify these expectations.

Following the same procedure as above, 500 conformations were generated for each of the four RGD-containing peptides and joint PCoorA projections were carried out. Since, in this case, bioactivity is clearly determined by the conformation (and stereochemistry) of the Arg−Gly−Asp region itself, the distance measure used to construct the joint distance matrix focused on the structure of this peptide. Thus distances between conformations were measured in dihedral angle space and summing over the seven dihedral angles (ϕ, ψ, ω) within this three amino acid region. This measure is suitable for comparing the four peptides to one another. Since the 500 conformations sampled for each peptide spanned a very broad range of energies (ranges of about 60 kcal/mol) the joint projection was very crowded. Much clearer PCA projections were obtained when the analysis was limited to the lowest 100 or 50 conformations from each peptide (energy ranges of about 10 kcal/mol). As before, the restriction of the calculation to the lower energy conformations is equivalent to a rough application of the Boltzmann factor. The results for both sample sizes were similar, but for clarity we report here the results obtained when using the 50 lowest conformations of each peptide.

Figure 9 shows the best 2D projection (first two principal axes) obtained when the four conformationally constrained RGD-containing analogues were jointly projected (diagonalization of a 200 × 200 matrix). The accuracy of this 2D projection is 50% (the accuracy of the associated 3D projection, not shown, is 67%). A schematic ellipse highlights the region of conformation space occupied by each of the peptides. A similar picture is retained when the projection is based on the 100 lowest energy conformations for each peptide (diagonalization of a 400 × 400 distance matrix), although the conformation space ellipsoids become broader.

The projection in Figure 9 clearly shows that the expected relationship between activity, specificity and conformation space holds for this group of conformationally constrained peptides. The two analogues which exhibit similar binding affinities and lack of specificity, GRGDSPC and G-dR-GDSPC, occupy the same area in conformation space, and have a similar conformation volume. The nonactive analogue, GRG-dD-SPC, on the other hand, occupies a separate region in conformation space. The fact that the conformation space of this analogue is disjoint from the two active analogues agrees well with the suggested scenario. Finally, the region of conformation space occupied by the VN-specific cyclic analogue is also separated from regions occupied by the other three molecules. This separation in conformation space follows the scenario suggested in Figure 1c. It seems that the cyclic analogue was able to focus right on top of the “VN region of bioactivity” and completely miss the “FN region of bioactivity”, hence its potency and specificity. This VN region of bioactivity is clearly broader than the area covered by the cyclic analogue itself, since the two nonselective peptides also show some affinity to this receptor.

Discussion

The present study suggests a conceptual framework for discussing and quantifying the effect of molecular flexibility and conformation constraints on the bioactivity of flexible molecules. It is well known that conformation constraints can dramatically alter the activity of numerous bioactive molecules (including, but not limited to, peptides). In fact, introduction of conformation constraints is one of the main avenues for optimizing candidate drug molecules (the other avenue is, of course, chemical modifications). However, while the role of chemical modifications is relatively well understood by today’s theory (electron transfer, ionic interactions, hydrophobic contacts, and so forth) a quantitative approach to the conformational aspect of the problem is lacking. This problem is especially evident in QSAR, where chemical properties and even 3D structural similarity are well accounted for, but for the most part molecular flexibility and conformation constraints elude quantification. In this study we quantifying these effects and show how they can be harnessed toward predictive ends.

The concept underlying this study is that conformation constraints and molecular flexibility are best discussed in terms of “conformation volumes”, rather than in terms of individual 3D structures. The idea is to shift the focus from 3D structures of a specific realization of the molecule to the rather abstract “occupied volume in conformation space”, V_{conf}, in which each conformation is no more than an abstract multidimensional point. This shift allows us to consider at once the whole world of conformations available to the molecule. Instead of asking whether one specific conformation is similar to another specific conformation, we ask the following questions: does the “world” of conformations accessible to a given molecule overlap with that of another molecule? how do conformation constraints shrink or shift the volume in conformation space accessible to the molecule? and so forth.

A second concept, introduced in this context, is the existence of an externally determined region of bioactivity R_{bio} in the ligand’s conformation space (the shaded areas in Figure 1). This region of bioactivity represents a collection of *all possible* ligand conformations which are compatible with the geometry of the host’s binding site, regardless of whether the ligand can actually adopt these conformations. Each host binding site, whether an enzyme’s or a receptor’s, is characterized by a specific geometry that presents a set of conformational requirements for the ligands to fulfill. Binding can occur when the ligand adopts a conformation compatible with the requirements presented by the host (assuming that the other, chemical, conditions are also obeyed). In terms of conformation spaces, this means that the conformational contribution to the binding affinity of a flexible ligand is related to the Boltzmann weighted overlap between two conformation space volumes: the host-prescribed region of bioactivity R_{bio} and the region in conformation space actually occupied by the ligand, V_{conf}. A large overlap between these two volumes indicates a high probability for the ligand to adopt a bioactive conformation, making it a highly potent binding agent. A small overlap, or lack of overlap altogether, indicates that the ligand is not active. The probability for this ligand to adopt a binding conformation is very small.

As discussed above, recent theoretical and methodological developments, which allow quantification and visualization of multidimensional conformation spaces, make these concepts tractable and suitable for practical application. In particular, these methods rely on principal component projections to reduce the dimensionality of large conformation samples. Jointly projecting several molecules onto the same low-dimensional (principal) space allows us to quantitatively compute the relative volume available to each molecule and quantify the degree of overlap between the region in conformation space accessible to one molecule and the region populated by another. Correlating this information with observed binding affinities allows us to infer about the host’s region of bioactivity.

The three scenarios for the relation between conformation constraints and bioactivity, which are discussed in this paper, are examples of the type of analysis offered by the above concepts. The fact that realistic systems actually follow these schemes is a very promising proof of concept. It is clear, however, that these three schemes cannot be exclusive and other schemes are likely to exist. For example, the dynamic situation upon which both ligand and receptor change their conformation during the binding process is not accounted for by any of the above scenarios.

The first scenario (Figure 1a) addresses the situation where the primary effect of conformation constraints, such as backbone cyclization, is to reduce the flexibility of the molecule. Reduced flexibility means that fewer conformations are available to the molecule, i.e., the region in conformation space accessible to it is smaller than the region accessible to the unconstrained analogue. This scenario was demonstrated by the series of alanine hexapeptide analogues. These exhibit a gradual reduction in the available conformation volume as the level of constraints increases from a single Ala to Pro substitution to a double Ala to Pro substitution and finally to a backbone cyclized analogue (88%, 55% and 12% respectively). In addition, this set of conformationally constrained analogues also demonstrates that the different constraints confine the molecule to different parts of its original conformation space.

The second scenario (Figure 1b) focuses on the case where the primary effect of some more delicate conformation constraints, such as L to D enantiomeric substitution in peptides, is to shift the region in conformation space accessible to the molecule relative to host-prescribed region of bioactivity (with little effect on the overall flexibility). As the accessible region of conformation space is shifted away from the region of bioactivity, the overlap between the two regions decreases and the bioactivity of the molecule decreases too. The validity of this scenario was demonstrated by a series of seven conformationally constrained substance P analogues (11 amino acid peptides), which showed a very strong correlation between overlaps in conformation space and the observed binding affinity. The quality of the correlation was such that it allowed us to use the multidimensional overlaps as a QSAR-type molecular descriptor, and predict binding affinities based on this descriptor. The resulting QSAR correlation was very strong. Recall, that in this case the other “chemical” parameters are unchanged (e.g., the chemical composition of the peptide), allowing us to focus on the pure effect of the conformation constraints. In a more general situation, such QSAR-type descriptors for the effect of conformation constraints will have to be weighted together with other “chemical” and “structural” parameters.

It should be stressed, that while our theory is formulated in terms of conformational overlaps with the host’s region of bioactivity, the application to substance P analogues involved a simplification. The conformation volume of the most active species was used in order to estimate the host's region of bioactivity. This simplification has the obviously serious drawback that it cannot identify molecules that are more active than the currently most active species, thus limiting its application. Therefore, this simplification should be considered only as a practical first step and not as the ultimate solution. For a general application of this approach we are currently developing methods for obtaining independent estimates of the host's region of bioactivity, which is based on sampling the conformation volume within the binding site.

Finally, the third suggested scenario (Figure 1c) focused on molecular specificity in the context of conformation space analysis. In principle, different receptors (or receptor subtypes) may have different conformational requirements from the ligand, manifested by different regions of bioactivity R_{bio}. A nonselective ligand would be flexible enough to engulf (or partially overlap) both regions of bioactivity. Selective ligands, on the other hand, preferentially correlate with one of these regions (nonactive ligands misses both regions altogether). The validity of this scenario was demonstrated by a family of four conformationally constrained RGD-containing septapeptides. Since there are at least two receptors that bind these peptides, it offers an opportunity to test the relationship between conformation spaces and binding specificity in a real chemical system. As predicted, the regions in conformation space occupied by this set of RGD-containing peptides were spatially separated according to the binding patterns. The highly specific peptide, the two active but nonspecific analogues as well as the nonactive analogue each occupy different regions in conformation space.

Conclusions

The present study offers both a conceptual and a practical framework for discussing and quantifying the effect of conformation constraints on the binding affinity of flexible molecules (assuming that the chemical composition of the molecules has not changed). These effects are formulated in terms of the overlap between the region in conformation space accessible to the ligand and the pre-prescribed conformational requirements imposed by the host molecule (region of bioactivity). The effect of conformation constraints on the ligand's flexibility and stereochemistry is either to shrink (reduce flexibility) or shift (change stereochemically) the volume in conformation space accessible to the ligand. This changes the overlap between this conformation volume and the host's region of bioactivity and the observed bioactivity changes along. Using a computational procedure for analyzing molecular conformation space, it was demonstrated that these concepts are valid in at least three different flexible molecular systems. Moreover, we have shown that the effects of conformation constraints in flexible molecules can be quantitatively accounted for and used in a QSAR context. Although the present study focused on bioactive peptides, it is not restricted to these type of systems. In principle, the suggested concepts and methodologies are applicable to the analysis of any flexible molecule in a broad range of binding and clustering situations.

Acknowledgment. We thank Alon Seri-Levy for stimulating discussions. This study was supported in part by grants from the Israeli Ministry of Science and from the Israeli Interuniversity Supercomputing Center.

References and Notes

(1) Pierschbacher, M. D.; Ruoslahti, E. *J. Biol. Chem.* **1987**, *262*, 17294−17298.

(2) Wang, J.; Dipasquale, A. J.; Bray, A. M.; Maeji, N. J.; Spellmeyer, D. C.; Geysen, H. M. *Int. J. Pept. Protein Res.* **1993**, *42*, 392−399.

(3) Horwell, D. C.; Howson, W.; Higginbottom, M.; Naylor, D.; Ratcliffe, G. S.; Williams, S. *J. Med. Chem.* **1995**, *38*, 4454−4462.

(4) Damewood, J. R., Jr. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1996; Vol. 9, pp 1−80.

(5) Lambert, M. H. In *Practical Applications of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Marcel Dekker: New York, 1997; pp 243−303.

(6) Tokarski, J. S.; Hopfinger, A. J. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792−811.

(7) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B. Q.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. *J. Am. Chem. Soc.* **1997**, *119*, 10509−10524.

(8) Gower, J. C. *Biometrika* **1966**, *53*, 325−338.

(9) García, A. n. E. *Phys. Rev. Lett.* **1992**, *68*, 2696−2699.

(10) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins* **1993**, *17*, 412−425.

(11) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567−2572.

(12) Case, D. A. *Curr. Opin. Struct. Biol.* **1994**, *4*, 285−290.

(13) Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Protein Sci.* **1998**, *7*, 649−666.

(14) Becker, O. M. *J. Comput. Chem.* **1998**, *19*, 1255−1267.

(15) Becker, O. M. *J. Mol. Struct. (THEOCHEM)* **1997**, *398−399*, 507−516.

(16) Levy, Y.; Becker, O. M. *J. Chem. Phys.*, submitted.

(17) Cascieri, M. A.; Huang, R.; Fong, T. M.; Cheung, A. H.; Sadowski, S.; Ber, E.; Strader, C. D. *Mol. Pharmacol.* **1992**, *41*, 1096−1099.

(18) Pierschbacher, M. D.; Rouslahti, E. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 5985−5988.

(19) Pytela, R.; Pierschbacher, M. D.; Ginsberg, M. H.; Plow, E. F.; Ruoslahti, E. *Science* **1986**, *231*, 1559−1562.

(20) Becker, O. M.; Karplus, M. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(21) Levy, Y.; Becker, O. M. *Phys. Rev. Lett.* **1998**, *81*, 1126−1129.

(22) Huber, T.; Torda, A. E.; van Gunsteren, W. F. *Biopolymers* **1996**, *39*, 103.

(23) Bruccoleri, R. E.; Karplus, M. *Bioploymers* **1990**, *29*, 1847−1862.

(24) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.

(25) MacKerell, A., et al. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(26) Becker, O. M. *Proteins* **1997**, *27*, 213−226.

(27) Free, S. M.; Wilson, J. W. *J. Med. Chem.* **1964**, *7*, 395−399.

(28) Eriksson, L.; Jonsson, J.; Hellberg, S.; Lindgren, F.; Skagerberg, B.; Sjostrom, M.; Wold, S. *Acta Chem. Scand.* **1990**, *44*, 50−56.